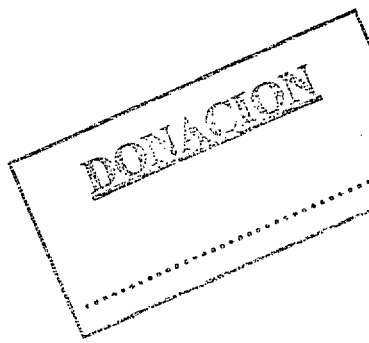
ANÁLISIS BAYESIANO DE SISTEMAS DE COLAS

TESIS DOCTORAL

Autora: María Concepción Ausín Olivera

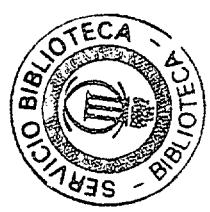
Directores: Rosa E. Lillo y Michael P. Wiper.



UNIVERSIDAD CARLOS III DE MADRID

Departamento de Estadística y Econometría

Getafe, diciembre de 2003



Dª. MARÍA CONCEPCIÓN AUSÍN OLIVERA, con D. N. I. 5653537 E

A U T O R I Z A:

A que su tesis doctoral con el título: ***"Análisis Bayesiano de Sistemas de Colas"*** pueda ser utilizada para fines de investigación por parte de la Universidad Carlos III de Madrid.

Leganés, 19 de febrero de 2004

Fdo.: María Concepción Ausín Olivera

A mi familia y a Pedro.

TRIBUNAL CALIFICADOR:

PRESIDENTE: Daniel Peña

VOCALES:

MARCEL F. NEUTS

DAVID RIOS INSUA

FABRIZIO RUGGERI

VOCAL SECRETARIO: CARMEN ARMERO CERVERA

CALIFICACIÓN: SOBRESALIENTE CUM LAUDE POR UNANIMIDAD

Leganés, 19 de febrero de 2004

Agradecimientos.

En el desarrollo de esta tesis, han intervenido, de un modo u otro, muchas personas a quienes quiero mostrar mi gratitud.

En primer lugar, quiero agradecer enormemente la dedicación de mis directores de tesis, Mike y Rosa. Ellos no han sido únicamente la guía y orientación en este trabajo, sino que se han implicado directamente constituyendo una ayuda fundamental en los problemas que han ido surgiendo en la elaboración de esta tesis. Muchas gracias por vuestro esfuerzo, por vuestra paciencia y por vuestra ayuda, no sólo desde el punto de vista profesional, sino también humano, aconsejándome a la hora de tomar decisiones, mostrando optimismo en los momentos más difíciles,... etc. Muchas gracias por haberme guiado en la perspectiva Bayesiana de la estadística, así como por haber convertido las dificultades de la Teoría de Colas en cuestiones asequibles con explicación intuitiva, por darme la oportunidad de conocer y aprender de diferentes inverstigadores a través de congresos, cursos y estancias, y por haber realizado una labor minuciosa de la revisión de todos los contenidos de esta tesis, así como de sus artículos derivados.

Agradezco al Departamento de Estadística y Econometría, al cual pertenezco, la organización del Doctorado en Ingeniería Matemática dentro del cual se ubica esta tesis. También, quiero dar las gracias a todos sus miembros y en especial, a mis compañeros de Leganés, que han sido un apoyo fundamental, ofreciéndome siempre su ayuda, tanto en la investigación como en la docencia, además de amigos con los que he disfrutado de muy buenos ratos en comidas y cafés. Gracias a María, Isaac, Ismael, José Luis, Javi, Teresa, Mounah, Vicente,... y muchos más, además de algunos que ya se han ido como Juan Carlos y Clara.

Quiero mostrar mi gratitud al Ministerio de Educación y Ciencia por el apoyo financiero ofrecido mediante el proyecto BEC 2000-0167 y a su investigador principal, Daniel Peña, por su disponibilidad.

Deseo manifestar mi agradecimiento a Fabrizio Ruggeri por haberme invitado a visitar el CNR, lo que constituyó una estancia muy fructífera cuyos resultados están contenidos en esta tesis, agradecerle su ayuda y sus consejos, así como la posibilidad de realizar el curso de verano *TED: Internet-based decision support* financiado por la European Science Foundation.

Mi familia y, en especial, mis padres y mis hermanas, me han ofrecido todo su cariño y su apoyo incondicional en la realización de esta tesis. Continuadamente, se han interesado por mí, preocupándose por mi trabajo, por mi estado de ánimo,... etc, alentándome en los momentos más difíciles y disfrutando conmigo de las buenas etapas. Muchas gracias de todo corazón, sin vosotros no hubiese sido posible llevar a cabo este trabajo. Quiero dar las gracias también a varios amigos que han influido de formas diferentes en esta tesis, a mis amigas del Poveda, a Mayte, Paqui, a mis compañeros de doctorado, a Dani, Montse y al grupo de la Autónoma y, por último, a Jesús Juan y a Raquel Díaz, por iniciarme hace ya unos años en este aventura.

Por último, no tengo palabras para Pedro. Él ha sido mi compañero y mi amigo día tras día. En el desarrollo de esta tesis, me he sentido muy afortunada porque además de haber estado cerca de mí en todo momento, ha sido un buen compañero de trabajo que me ha ayudado con muchos temas de la tesis. Además, como hemos pasado por las mismas situaciones, nos ha sido más fácil comprender y compartir las dificultades y los buenos momentos. Muchas gracias por tu cariño, tu comprensión y tu trabajo.

Índice general

Índice de figuras

Índice de tablas

Introducción y resumen.

Las aplicaciones del análisis de sistemas de colas constituyen todo un abanico de posibilidades que abarca desde los problemas más tradicionales, como los relacionados con el transporte o la gestión sanitaria, hasta cuestiones más recientes que surgen en el tráfico de datos entre ordenadores o en redes de telecomunicaciones. En este último caso, se sitúan, por ejemplo, las redes de datos conmutadas (*data switching networks*), que incluyen a las redes de telefonía básica, en las que los paquetes de datos (*data packets*) deben permanecer en áreas de almacenaje temporal (*buffers*) de los nodos de conmutación (*switching nodes*) y formar una línea de espera hasta que se produzca su transmisión, véase, por ejemplo, Bertsekas y Gallager (1992).

La Teoría de Colas clásica es la disciplina que aborda el análisis, desde un punto de vista probabilístico, del comportamiento de las llegadas, el servicio, la espera y otras cuestiones asociadas a un sistema de colas haciendo uso de los conceptos y métodos del campo de los procesos estocásticos. Esta teoría se fundamenta en modelos estocásticos que describen el comportamiento aleatorio de un sistema de colas con la finalidad de formular predicciones sobre sus tiempos de espera, el número de clientes en un instante dado, la longitud de un periodo activo..., y con el objetivo de obtener diseños y control óptimo de estos sistemas desde el punto de vista de la optimización de costes. La Teoría de Colas puede considerarse una rama de la Matemática Aplicada, de la Investigación Operativa o de los Procesos Estocásticos, sin embargo desde el trabajo de Erlang (1909), considerado el pionero en el análisis de colas, hasta la actualidad, el desarrollo de esta materia ha sido enormemente fructífero hasta convertirse en una disciplina independiente que incluye una multitud de libros, artículos, congresos y revistas destinados a su investigación.

En contraste con la modelización estocástica de los sistemas de colas, el estudio estadístico de los mismos ha sido comparativamente mucho más escaso. Tradicionalmente, en la Teoría de Colas clásica se ha asumido, en general, que el comportamiento de las llegadas y los servicios en un sistemas de colas, así como los parámetros que lo determinan son conocidos, centrando el estudio de los mismos en aspectos descriptivos asociados a su funcionamiento. Sin embargo, en la práctica, la única información útil de la que se dispone en este sentido se basa, por un lado, en la experiencia de los especialistas conocedores del funcionamiento del sistema y por otro, en conjuntos de datos procedentes de experimentos para la observación del mismo. Consecuentemente, es evidente la importancia del empleo de herramientas estadísticas, no sólo con la finalidad de obtener estimaciones de los parámetros implicados en un modelo de colas elegido, sino también con el objeto de desarrollar procedimientos que permitan la selección de un modelo de colas adecuado para describir el sistema observado, así como la predicción de diferentes cantidades de interés en el mismo, tales como el número de clientes o el tiempo de espera en cola.

Cuando se aborda el análisis estadístico de un sistema de colas, sea cual sea la metodología empleada, la recogida de datos se centra, generalmente, en la observación de la conducta de las llegadas y los servicios en el sistema. El motivo de esta elección se basa principalmente en la posibilidad de integrar esta información en los diferentes resultados clásicos de la Teoría de Colas, fundamentados en el proceso de llegadas y de servicio que determinan el modelo de colas (junto con otras características fáciles de determinar tales como el número y disposición de los servidores o el número total de clientes que el sistema puede admitir). De este modo, el objetivo final es la predicción de las cantidades que realmente interesan en el sistema como son

los tiempos de espera en cola. La metodología Bayesiana permite un procedimiento sencillo de incorporar la incertidumbre resultante de las estimaciones asociadas al proceso de llegadas y de servicio en la predicción de estas cantidades de interés. Además de esta característica, la perspectiva Bayesiana ofrece muchas otras cualidades para el análisis de sistemas de colas que se comentarán más adelante, tales como la obtención directa de probabilidades esenciales en el sistema o la incorporación de restricciones en el espacio paramétrico.

El análisis Bayesiano de sistemas de colas es actualmente un área de investigación bastante desarrollada. Sin embargo, en la mayoría de los trabajos, se han considerado modelos de colas Markovianos para la inferencia y la predicción de los sistemas observados y, en muchas ocasiones, estos modelos no son apropiados para describir situaciones reales. Por ejemplo, en problemas de transmisión de voz, la duración de una llamada sí que suele distribuirse exponencialmente, pero en problemas de transmisión de paquetes de datos digitales, la longitud de los mismos puede ser fija o tener unos límites impuestos por diferentes restricciones de carácter físico. Esta situación motiva el interés de esta tesis que consiste, fundamentalmente, en el análisis Bayesiano de sistemas de colas generales que no asuman una estructura rígida para el comportamiento de las llegadas y de los servicios al sistema, de modo que sea posible su descripción mediante un modelo de colas que se adecúe a los datos recogidos a partir de la observación del sistema.

En esta tesis, se proponen procedimientos Bayesianos para la inferencia, predicción y diseño de diferentes sistemas de colas que incluyen modelos con distribuciones generales para el tiempo entre las llegadas y/o el tiempo de servicio, con uno o varios servidores y con capacidad finita e infinita. Con las observaciones tomadas del proceso de llegadas y de servicio en el sistema, en primer lugar, se desarrollan métodos de estimación Bayesiana de densidades que permiten aproximar las distribuciones desconocidas asociadas a estos procesos. La estimación está basada en mixturas de distribuciones con un número desconocido de componentes que se aborda con diferentes métodos MCMC de dimensión paramétrica variable. Los modelos de mixtura seleccionados son de tipo PH, lo cual ofrece muy buenas propiedades para su posterior aplicación en colas. A continuación, se analiza la congestión del tráfico en cada sistema observado basándose en la inferencia desarrollada sobre los parámetros del mismo. En caso de que el sistema sea estable, se describen procedimientos para la estimación Monte Carlo de las medidas de interés de cada modelo en equilibrio, tales como las distribuciones predictivas del número de clientes en el sistema, del tiempo de espera en cola o de la longitud de los periodos de ocupación. La predicción Bayesiana desarrollada se basa en diferentes propiedades de los sistemas en los que intervienen distribuciones PH y en algunos resultados conocidos de la Teoría de Colas clásica. En los modelos de colas con varios servidores, se abordan también problemas de diseño en los que el objetivo consiste en decidir el número óptimo de servidores que minimizan una función de coste que depende de las distribuciones estacionarias estimadas. La metodología desarrollada se ilustra con datos reales procedentes de un hospital geriátrico de Londres y un establecimiento bancario de Madrid. Por último, se analiza el comportamiento transitorio y el periodo de ocupación de modelos de colas completamente generales con un único servidor. Se obtiene un resultado que permite extraer numéricamente las raíces complejas de unas ecuaciones implicadas en las distribuciones transitorias de interés.

Las aportaciones de esta tesis no se limitan a los problemas de colas y tienen aplicaciones interesantes fuera de este contexto. Por un lado, los algoritmos propuestos para la estimación Bayesiana de densidades pueden utilizarse para la aproximación de variables continuas y positivas que se requieren en otras materias como, el Análisis de Supervivencia o la Teoría del Riesgo. Además, en estas mismas disciplinas existen también resultados basados en aproximaciones de tipo PH que pueden incorporarse siguiendo la metodología propuesta en esta tesis. Por último, el análisis estadístico de las funciones de coste considerado en esta tesis puede adaptarse a diferentes estructuras de coste que se formulan en otras áreas de investigación, como los sistemas de inventariado, donde frecuentemente se suponen conocidos los parámetros o se utilizan estimaciones puntuales para los mismos.

A continuación, se resume brevemente el contenido de cada uno de los Capítulos de la tesis.

En el Capítulo 1, se revisan cuestiones generales relacionadas con la Teoría de Colas clásica y con el análisis estadístico de los sistemas de colas. Por un lado, se exponen los conceptos básicos de esta teoría y se describen

brevemente las características de diferentes modelos de colas, partiendo de los sistemas Markovianos más sencillos, hasta modelos más generales que requieren el uso de técnicas más complejas para su análisis, como los métodos matriciales. Se incluyen también algunos resultados clásicos que se utilizarán a lo largo de esta tesis. Por otra parte, se exponen los objetivos y las dificultades en el análisis estadístico de los sistemas de colas, mencionando algunos procedimientos clásicos para la inferencia, así como una extensa recapitulación de los análisis Bayesianos más relevantes existentes en la literatura para sistemas de colas. Se incluye también una breve descripción del procedimiento para la inferencia y predicción Bayesiana en el modelo de colas más sencillo, el sistema M/M/1.

El Capítulo 2 está destinado al análisis de los procesos de llegadas y de servicio en un sistema de colas general. Concretamente, se proponen cuatro algoritmos diferentes para la estimación Bayesiana de la distribución de una variable aleatoria desconocida definida en la semirecta real positiva. Estos métodos se utilizarán en los Capítulos siguientes para aproximar las distribuciones desconocidas del tiempo entre las llegadas y del tiempo de servicio en un modelo de colas general. Los procedimientos de estimación de densidades están basados en modelos semiparamétricos consistentes en mixturas de distribuciones con la cualidad de ser de tipo PH, lo cual implicará propiedades muy buenas para su aplicación en modelos de colas en Capítulos posteriores. Como es habitual, para desarrollar inferencia Bayesiana, se hace uso de los métodos basados en Cadenas de Markov Monte Carlo (MCMC), de los que se incluye una breve introducción. Además, puesto que se asume que el número de términos que intervienen en los modelos de mixtura considerados es desconocido, los algoritmos propuestos se basan en técnicas MCMC de dimensión paramétrica variable, como son los métodos de salto reversible o los algoritmos MCMC en tiempo continuo. Los procedimientos propuestos se comparan y se ilustran con datos simulados y reales.

En el Capítulo 3, se desarrolla inferencia y predicción Bayesiana para el modelo de colas M/G/1 en equilibrio. Dados los datos recogidos del proceso y de llegadas al sistema, se describen procedimientos Bayesianos para aproximar las distribuciones predictivas del número de clientes presentes en el sistema, del tiempo de espera en cola y de la longitud de los periodos de ocupación. La metodología desarrollada requiere, por un lado, de los procedimientos para la inferencia sobre el tiempo general de servicio obtenidos en el Capítulo 2, y por otra parte, de algunos de los resultados conocidos en la Teoría de Colas sobre los modelos M/PH/1. Al inicio de este Capítulo, se incluye una introducción a las distribuciones PH, así como la relación de propiedades necesarias relativas a los sistemas M/PH/1, que se obtienen en Neuts (1977, 1981). La integración de estos resultados para la predicción Bayesiana es posible gracias a que los modelos de mixtura considerados para la aproximación del tiempo de servicio son de tipo PH. Además, como se comentó anteriormente, el método Bayesiano permite la incorporación natural del error de estimación del tiempo de servicio en las distribuciones predictivas de las medidas mencionadas. Los procedimientos se ilustran sobre sistemas de colas cuyos tiempos de servicio son los conjuntos de observaciones incluidos en el Capítulo 2.

En el Capítulo 4, se proponen procedimientos de estimación Bayesiana para sistemas de colas con varios servidores. Se plantean también problemas de diseño óptimo en los que el número de servidores es la variable de control. La ilustración del análisis se aplica sobre dos situaciones reales concretas que se ubican en un hospital geriátrico de Londres y en un establecimiento bancario de Madrid, las cuales pueden describirse mediante los sistemas M/G/c/c y GI/M/c, respectivamente. En ambos modelos de colas, se desarrollan métodos para la predicción Bayesiana de las medidas de interés en equilibrio que permiten estimar cantidades como la distribución del número de camas ocupadas en el hospital o del tiempo de espera en la sucursal bancaria, a partir de los datos reales de los que se dispone. Además, en este Capítulo, se formulan para cada sistema, funciones de coste con el fin de abordar el diseño del mismo y decidir, en cada caso, el número óptimo de servidores. Los costes a los que se somete el sistema se evalúan en el estado estacionario de modo que las funciones de coste dependen de las distribuciones estacionarias de las cantidades de interés estimadas previamente. Aunque la metodología desarrollada se ilustra sobre la decisión del número de camas en el hospital y del número de servidores en el banco, se muestra cómo aplicarla para el diseño de cualquier sistema de colas general, GI/G/c/K, con $K \leq \infty$.

El Capítulo 5 se dedica al análisis Bayesiano del modelo de colas general GI/G/1. A diferencia de los

Capítulos anteriores, se estima no sólo el comportamiento del sistema en equilibrio, sino también en sus estados transitorios. Concretamente, se describe cómo aproximar las distribuciones predictivas del número de clientes en el sistema en cualquier instante de tiempo, τ, así como la distribución predictiva del tiempo de espera en cola de un cliente que llega al sistema en cualquier instante, τ. Además, se obtiene una estimación de la distribución de la longitud del periodo de ocupación en el sistema. Para desarrollar estos procedimientos se utilizan unos resultados de la Teoría de Colas clásica que fueron obtenidos por Bertsimas y Nakazato (1992). La idea de Bertsimas y Nakazato (1992) se basa en el método de las fases y consiste en la aproximación de la distribución general del tiempo entre las llegadas y de servicio del sistema GI/G/1 con uno de los modelos de mixtura considerados en el Capítulo 2. Este método requiere, entre otros cálculos, la extracción de raíces de ecuaciones complejas que se sugieren obtener de forma simbólica. Puesto que el cálculo simbólico no es compatible con los métodos de estimación MCMC, en este Capítulo, se demuestra un resultado a partir del cual se deduce un procedimiento numérico para la extracción de las raíces mencionadas.

Por último, en el Capítulo 6, se exponen comentarios y extensiones generales y naturales de la tesis. Aparte, se añaden comentarios y extensiones individuales al final de cada Capítulo.

Algunos de los contenidos de esta tesis han sido publicados en Ausín et al. (2003) y en Ausín et al. (2004)

Capítulo 1

Sistemas de colas e inferencia Bayesiana.

Este es un Capítulo introductorio que consta de dos partes diferenciadas y que pretende incorporar nociones básicas de los dos temas que trata esta tesis. Por un lado, se ofrece una breve introducción a la Teoría de Colas, así como algunas cuestiones preliminares relacionadas con los modelos de colas generales que se consideran en esta tesis. Por otro lado, se presentan las ideas básicas del análisis estadístico de sistemas de colas con una revisión de la literatura en esta materia.

Este Capítulo está dividido en cuatro Secciones. En la Sección 1.1, se incluye una breve descripión de los elementos básicos que caracterizan a un sistema de colas, así como sus medidas de interés, tales como el tamaño del sistema o el tiempo de espera en cola. En la Sección 1.2, se ofrece una visión general de los modelos de colas Markovianos y se presentan las medidas asociadas al caso particular más elemental, que es el sistema M/M/1. La Sección 1.3 está destinada a los modelos de colas no Markovianos, donde se exponen, en primer lugar, una breve descripción de la definición y propiedades de los sistemas M/G/1 y GI/M/c. A continuación, una introducción muy resumida a las distribuciones PH y los métodos matriciales de Neuts (1981). Y por último, se comentan, entre otras, las ideas de Bertsimas y Nakazato (1992) para el análisis del comportamiento transitorio del sistema GI/G/1. En la Sección 1.4, se considera el análisis estadístico de los sistemas de colas, donde se exponen los objetivos y dificultades que se plantean para este estudio. A continuación, se citan algunos trabajos para la inferencia clásica de sistemas de colas, así como una extensa revisión de los análisis Bayesianos, que incluye una breve descripción para la inferencia y predicción Bayesiana del sistema M/M/1.

1.1. Fundamentos de la Teoría de Colas.

En esta Sección, se introducen los elementos básicos que describen un sistema de colas, así como la notación empleada habitualmente para identificar los diferentes modelos. Se incluye también la descripción de las medidas más interesantes asociadas a un sistema de colas y algunas de sus propiedades más generales.

1.1.1. Caracterización de un sistema de colas.

Un sistema de colas es una estructura en la que se producen llegadas en el tiempo de ciertas unidades denominadas *clientes* para recibir algún tipo de operación o servicio por parte de otras unidades que se denominan *servidores*. El mecanismo de un sistema de colas es el siguiente. En el instante en el que se

produce la llegada de un cliente, éste comienza a ser atendido si existe algún servidor desocupado. Si por el contrario, todos los servidores se encuentran ocupados atendiendo a otros clientes, el cliente que llega debe aguardar para recibir su servicio formando una línea de espera o bien, abandonar el sistema, según estén definidas las caraterísticas del mismo.

Los términos genéricos, *clientes* y *servidores*, no se refieren exclusivamente a personas físicas. Por ejemplo, en el caso de las redes de datos conmutadas citadas anteriormente en la Introducción, los clientes son los paquetes de datos que llegan a un nodo de conmutación y los servidores son los canales de transmisión. Otro ejemplo se encuentra en problemas de asignación de trabajos a diferentes unidades centrales de proceso (*CPU job scheduling problems*) en los que los clientes son los procesos computacionales a desarrollar y los servidores son los CPU o dispositivos I/O.

En la mayoría de las situaciones, se consigue una descripción adecuada del funcionamiento del sistema con cinco características básicas que se pueden sintetizar de forma compacta siguiendo la notación de Kendall (1953) como $A/S/c/K/R$. La intepretación es la siguiente: A y S reflejan la conducta de las llegadas y los servicios, respectivamente, c es el número de servidores, K representa la capacidad del sistema y, por último, R es la disciplina establecida para acceder al servicio. A continuación se describen estos cinco elementos más detalladamente.

Los procedimientos que determinan el comportamiento de las llegadas de los clientes al sistema y de su servicio pueden ser deterministas o aleatorios. En el primer caso, las llegadas se producen a intervalos fijos de tiempo y la duración del servicio es constante. Sin embargo, en la mayoría de las situaciones, las llegadas y/o los servicios suceden de forma aleatoria y, en estos casos, se hace uso de los procesos estocásticos para su descripción probabilística. Concretamente, en esta tesis, se supone que los procesos de llegadas y de servicio están constituidos por sucesiones independientes de variables aleatorias i.i.d. De este modo, ambos procesos quedan determinados por las distribuciones de las variables aleatorias que describen el tiempo entre llegadas, que se denotará por A, y el tiempo de servicio, S. Como es habitual en la Teoría de Colas, si alguna de las variables A o S se distribuye exponencialmente se denota con M, que hace referencia a que el proceso de llegadas o de servicio sigue un proceso de Markov. Si la variable es degenerada y, por tanto, determinista, se indica con la letra D. Si, por el contrario, está regida por una distribución determinada se simboliza con su correspondiente abreviatura, por ejemplo, una distribución Erlang mediante Er o una mixtura de exponenciales (o hiperexponencial) con H_k. Por último, si la distribución es general y desconocida se denota por G y, en particular, si se trata del proceso de llegadas se enfatiza la indepedencia entre las llegadas con la notación GI.

Existen también otros aspectos necesarios para describir la conducta de las llegadas y los servicios. Por ejemplo, las llegadas se pueden producir en grupos, o análogamente, existen situaciones en las que se atiende simultáneamente a un grupo de clientes. Otro factor a tener en cuenta es la denominada dependencia del estado del sistema que aparece, por ejemplo, en casos en los que los clientes deciden no incorporarse a la cola de espera por ser demasiado extensa o en los que tras un tiempo de espera en el sistema deciden abandonarlo. Estos fenómenos reciben el nombre de impaciencia de los clientes. El proceso de servicio puede ser también dependiente del estado del sistema si, por ejemplo, su funcionamiento es más o menos rápido según el número de clientes esperando en cola. Por último, el proceso de llegadas y/o el de servicio puede ser independiente o no del tiempo, lo que da lugar a un proceso estacionario o no estacionario, respectivamente. A lo largo de esta tesis, se asume que las llegadas y los servicios se producen individualmente y que ambos procesos son independientes entre sí, independientes del estado del sistema e independientes del tiempo.

Por otro lado, los c servidores que atienden las llegadas de los clientes al sistema pueden estar dispuestos en paralelo o en serie. Además, los clientes pueden esperar formando una única línea de espera o formando una cola para cada servidor. En esta tesis, se asume que los servidores se encuentran en paralelo y que hay una sóla línea de espera, como se ilustra en el ejemplo de la Figura 1.1.

La capacidad del sistema de colas, K, es el número total de clientes que el sistema puede admitir. Si la capacidad es finita, $K < \infty$, existe una limitación en la línea de espera de modo que si el número de clientes

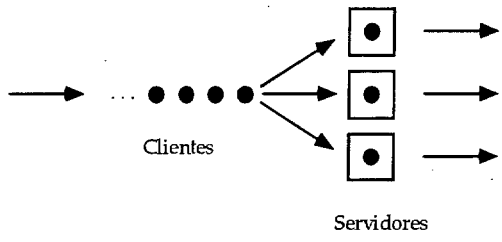


Figura 1.1: Ilustración de un sistema de colas con 3 servidores en paralelo y una única línea de espera.

presentes en el sistema es igual a K, no se admite la incorporación de ningún cliente nuevo hasta que no se libere espacio en la cola. Como es natural, si la capacidad es infinita la línea de espera no está acotada. En esta tesis, se consideran diferentes sistemas con uno y con varios servidores, con capacidad finita e infinita y además, un caso particular de sistemas con capacidad finita, los denominados sistemas de pérdida, en los que la capacidad finita del sistema coicide con el número de servidores.

Por último, la disciplina del sistema es el procedimiento por el cual se seleccionan los clientes esperando en cola para acceder al servicio. Los modelos considerados en esta tesis están regidos por la disciplina FIFO (*First In First Out*) que se caracteriza porque los clientes son atendidos en el orden de llegada. Esta es la disciplina más frecuente aunque existen otras en las que el primer cliente que se atiende es el último que ha llegado, o en las que existen distintos tipos de clientes con distintas prioridades.

Nótese que, si al denotar un determinado modelo de colas no se indica ninguna información sobre la capacidad y/o la disciplina, por ejemplo, en un sistema M/M/c, se sobreentiende, por defecto, que la capacidad es infinita y la disciplina FIFO.

1.1.2. Medidas de interés en un sistema de colas y su condición de ergodicidad.

El objetivo fundamental en el análisis de un modelo de colas no son los procesos de llegadas y de servicio, sino la descripción del comportamiento probabilístico de otras cantidades relacionadas con la espera y acumulación de clientes en el sistema y con la ocupación de los servidores. No obstante, estas medidas se ven influenciadas directamente por los procesos de llegadas y de servicio. Se definen a continuación algunas de estas cantidades de interés,

$$N(\tau) \;:\; \text{Número de clientes en el sistema en el instante } \tau. \tag{1.1}$$
$$N_q(\tau) \;:\; \text{Número de clientes esperando en cola en el instante } \tau. \tag{1.2}$$
$$N_b(\tau) \;:\; \text{Número de servidores ocupados en el instante } \tau. \tag{1.3}$$
$$W(\tau) \;:\; \text{Tiempo de espera en cola de un cliente que llega al sistema en el instante } \tau. \tag{1.4}$$

Obsérvese que, en cualquier instante, el número total de clientes en el sistema es igual al número de clientes esperando en la línea de espera, o cola, más el número de clientes en el servicio, que equivale al número de servidores ocupados,

$$N(\tau) = N_q(\tau) + N_b(\tau),$$

y el tiempo de espera en cola, $W(\tau)$, se inicia en el instante, τ, en el que se supone la llegada de un nuevo cliente y finaliza cuando éste accede al servicio. Todas estas cantidades son procesos estocásticos y por tanto, en cada instante, τ, cada una de ellas constituye una variable aleatoria cuya distribución es muy difícil de obtener en la mayoría de los casos. Sin embargo, en algunas ocasiones, es posible obtener una descripción de estas propiedades del sistema si se alcanza el equilibrio, lo cual sucede cuando el sistema ha permanecido

en funcionamiento durante un largo periodo de tiempo y si se verifican unas condiciones que se describen a continuación.

En muchas situaciones reales, es importante conocer la utilización de los recursos de los que se dispone en el sistema de colas que se pretende analizar, es decir, se desea conocer cuál es su grado de ocupación. Una medida de esta congestión en un modelo de colas general, GI/G/c/K, con $K \leq \infty$, es la *intensidad de tráfico*, ρ, que se define como,

$$\rho = \frac{\text{tasa media de llegadas}}{\text{tasa media de servicio}} = \frac{E[S]}{E[A]}, \tag{1.5}$$

siendo $E[S]$ y $E[A]$ el valor esperado de la distribución de los tiempos de servicio y de los tiempos entre llegadas de clientes, respectivamente. Obsérvese que, si $\rho > c$, entonces el número medio de llegadas al sistema (tasa media de llegadas) es mayor que el número medio máximo de clientes que el sistema puede atender (c veces la tasa media de servicio). Consecuentemente, en esta situación, la cola de espera será cada vez más larga a medida que pase el tiempo, siempre que la capacidad del sistema sea infinita, $K = \infty$. Bajo estas condiciones, el tamaño del sistema, $N(\tau)$, tenderá a infinito a medida que aumente el valor de τ y por tanto, no existirá su distribución estacionaria o en equilibrio. Nótese también que, si $\rho = c$, tampoco se alcanza el equilibrio, a no ser que se trate de sistemas con servicios y llegadas deterministas, es decir, modelos de colas D/D. Con la excepción de estos sistemas, para el resto de modelos GI/G/c, la condición que asegura la existencia de la distribución en equilibrio del sistema, que es la distribución de probabilidad de los estados de la cola en el límite, es,

$$\text{Condición de equilibrio: } \rho < c. \tag{1.6}$$

Si se verifica esta condición se dice que el sistema es estable y entonces, existen las distribuciones en equilibrio de todas las medidas asociadas al sistema. Por tanto, las denominadas distribuciones transitorias, $N(\tau)$, $N_q(\tau)$ y $W(\tau)$, convergen a sus distribuciones estacionarias, N, N_q y W, respectivamente, cuando τ tiende a infinito, es decir, por ejemplo,

$$P(N = n) = \lim_{\tau \to \infty} P(N(\tau) = n).$$

Obsérvese que, siempre que el número de servidores, c, sea finito, no se requiere que el sistema sea estable para que exista la distribución estacionaria, N_b, de $N_b(\tau)$, puesto que su valor está acotado en c. Por otro lado, nótese también que en los sistemas con un número infinito de servidores, siempre se verifica la condición de equilibrio, (1.6).

Se indican a continuación dos resultados generales que se utilizarán en esta tesis, véase Gross y Harris (1985) para más resultados de este tipo. En cualquier modelo de colas GI/G/c, si se verifica la condición de equilibrio, (1.6), la probabilidad estacionaria de que el sistema de colas esté vacío en un instante aleatorio viene dada por,

$$P(N = 0) = 1 - \rho, \tag{1.7}$$

y además, bajo esta condición, el número medio de servidores ocupados es,

$$E[N_b] = \rho. \tag{1.8}$$

Intuitivamente, se pueden interpretar estas afirmaciones puesto que la intensidad de tráfico, ρ, (1.5), es una medida de la ocupación del sistema.

La condición (1.6) se denomina también condición de ergodicidad, lo cual hace referencia a la ergodicidad del proceso asociado a los estados del sistema. Básicamente, el proceso $N(\tau)$ es ergódico si sus características se pueden determinar a partir de una única realización del proceso. Si un proceso es ergódico, se convierte en un proceso estacionario cuando el tiempo, τ, tiende a infinito. Formalmente, el proceso $N(\tau)$ es ergódico si se verifica que,

$$E[N(T)] = \int_0^T N(\tau)\, d\tau.$$

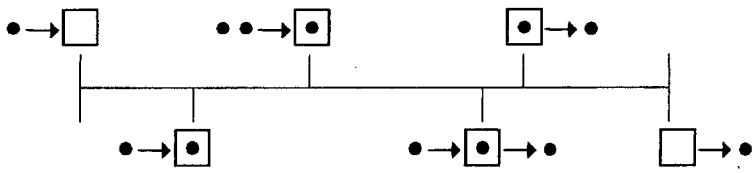


Figura 1.2: Ilustración esquemática de la longitud del periodo de ocupación de un sistema con un único servidor, representado con un cuadrado, y donde los clientes se indican con un punto negro.

Intuitivamente, un proceso es ergódico si la media del tiempo de recurrencia a cada estado es finita.

Es importante señalar que, aunque frecuentemente el análisis de un sistema de colas se centra en sus propiedades en equilibrio, en muchas ocasiones es importante también conocer el comportamiento transitorio del mismo. Por ejemplo, es interesante examinar la distribución de probabilidad del número de clientes en el sistema, $N(\tau)$, en cada instante τ, en lugar de considerarlo únicamente en el estado estacionario, y conocer, de este modo, la evolución temporal del modelo de colas. Sin embargo, en la mayoría de los casos es muy complicado obtener este tipo de distribuciones y los escasos resultados de los que se dispone en este sentido no tienen, generalmente, un tratamiento sencillo. En el Capítulo 5 de esta tesis, se aborda, entre otras cuestiones, el problema de la estimación del comportamiento transitorio de un modelo de colas general.

Además de las cantidades mencionadas hasta ahora, existen otras medidas de interés considerable en un sistema de colas que son la longitud de los periodos de ocupación y de desocupación del sistema, denotados por B e I, respectivamente. Un periodo de ocupación se define como el tiempo que transcurre entre la llegada de un cliente a un sistema vacío y el momento siguiente en el que el sistema se queda de nuevo vacío tras el abandono de un cliente servido. Esta definición se ilustra de modo esquemático en la Figura 1.2. Como es natural, el periodo de desocupación se define como el tiempo que transcurre desde el instante en el que un cliente servido abandona el sistema dejándolo vacío hasta la llegada de un nuevo cliente.

1.2. Sistemas de colas Markovianos.

En un modelo de colas Markoviano es posible describir la dinámica propia de los diferentes estados en los que se puede encontrar el sistema mediante una cadena de Markov de parámetro continuo. La ventaja fundamental de los sistemas que verifican esta propiedad es que frecuentemente se puede obtener la distribución de probabilidad de los estados de la cola en el límite, o sea, su distribución estacionaria, a partir de unas ecuaciones diferenciales, conocidas como ecuaciones de Kolmogorov, asociadas a estos procesos, véase, por ejemplo, Gross y Harris (1985) para una información detallada sobre esta materia.

Los modelos de colas Markovianos más simples son aquellos en los que el proceso asociado al número de clientes presentes en el sistema, $N(\tau)$, (1.1), se puede describir mediante un proceso de nacimiento y muerte (*birth-death process*), que es un caso particular de cadena de Markov de parámetro continuo. En estos casos, la llegada de un cliente se interpreta como un nacimiento y el abandono del sistema por un cliente servido como una muerte. En esta clase de modelos, se engloban todos los sistemas en los que las llegadas se producen individualmente según un proceso de Poisson y en los que el tiempo de servicio se distribuye exponencialmente atendiendo a los clientes de uno en uno, es decir, los sistemas de colas M/M. En estos modelos, la distribución en equilibrio del tamaño del sistema, N, se obtiene como la solución estacionaria de las ecuaciones de Kolmogorov asociadas a un proceso de nacimiento y muerte cuyas tasas dependen del modelo de colas que se considere. La distribución estacionaria de otras caracterísicas, como el tiempo de espera en cola, W, se obtiene frecuentemente a partir de la distribución de N. En la siguiente Subsección, se exponen los resultados más relevantes del modelo de colas más sencillo, el sistema M/M/1.

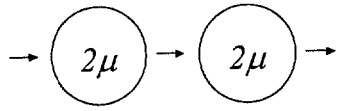


Figura 1.3: Representación de una distribución Erlang con dos fases exponenciales.

Existen otros modelos de colas Markovianos, que no son de tipo nacimiento-muerte, basados en cadenas de Markov de parámetro continuo en las que se permite la transición de un estado a otro que difiera en más de una unidad. Un ejemplo son los sistemas de colas $M^X/M/1$ y $M/M^X/1$, con llegadas o servicios en grupos, respectivamente. En estos casos, la solución estacionaria de las correspondientes ecuaciones de Kolmogorov, que es la distribución en equilibrio del tamaño del sistema, N, se obtiene con procedimientos basados en su función generatriz de probabilidad, véase, por ejemplo, Gross y Harris (1985).

Otros ejemplos de sistemas Markovianos más generales son los modelos M/Er/1 y Er/M/1, con distribución Erlang para el el tiempo de servicio o el tiempo entre las llegadas, respectivamente. La distribución Erlang constituye el primer modelo de distribución basado en el método de las etapas (*method of stages*), que se considerará frecuentemente, en esta tesis, para otros modelos de distribución más generales. La idea consiste en descomponer el tiempo entre las llegadas o el de servicio en una serie de fases exponenciales con el objeto de describir situaciones en las que el modelo de distribución exponencial no es apropiado. La Figura 1.3 ilustra esquemáticamente la estructura de una distribución Erlang con dos etapas. Su definición y propiedades se detallarán más adelante, en los Capítulos 2 y 3. La distribución Erlang permite conservar la propiedad Markoviana en procesos que describen los sistemas M/Er/1 y Er/M/1, pero, en estos casos, los estados de la cola no indican el número de clientes, sino el número de fases de servicio o de llegadas, respectivamente, en el sistema. Las soluciones de las correspondientes ecuaciones de equilibrio se obtienen nuevamente en términos de funciones generadoras de probabilidad, véase, por ejemplo, Kleinrock (1975).

Es importante señalar que existe una dualidad entre los sistemas M/Er/1 y Er/M/1 y los modelos $M^X/M/1$ y $M/M^X/1$. Nótese que el sistema M/Er/1 también se puede interpretar considerando las X fases del tiempo de servicio Erlang como la llegada de X clientes en grupo. De este modo, se observa que la cadena que describe el número de fases en el sistema M/Er/1 es la misma que describe el número de clientes en el sistema $M^X/M/1$. Análogamente, esta dualidad se produce entre los sistemas Er/M/1 y $M/M^X/1$.

Globalmente, lo importante de estos sistemas es que el tener propiedades Markovianas permite obtener expresiones analíticas explícitas de las cantidades de interés.

1.2.1. El sistema de colas M/M/1

El modelo de colas M/M/1 constituye el sistema básico en la Teoría de Colas. En este modelo, los clientes llegan al sistema siguiendo un proceso de Poisson de tasa λ y son atendidos por un único servidor con tiempos de servicio i.i.d. según una exponencial de tasa μ. Con la notación introducida en la Subsección 1.1.1, la capacidad de este sistema es infinita, la disciplina FIFO y las llegadas y los servicios se producen individualmente y son independientes entre sí, independientes del estado del sistema e independientes del tiempo.

En el sistema M/M/1, la intensidad de tráfico, definida en (1.5), viene dada por,

$$\rho = \frac{\lambda}{\mu}, \tag{1.9}$$

y su condición de equilibrio, introducida en (1.6), es $\rho < 1$. Como se ha mencionado anteriormente, el proceso, $N(\tau)$, (1.1), que describe el número de clientes presentes en el sistema M/M/1 es un proceso de nacimiento

y muerte cuya solución estacionaria existe si se verifica la condición de equilibrio, $\rho < 1$, y viene dada por, véase, por ejemplo, Gross y Harris (1985),

$$P(N = n) = (1 - \rho)\,\rho^n, \qquad n = 0, 1, ... \tag{1.10}$$

que es una distribución geométrica de parámetro ρ. Obsérvese que la probabilidad de que el sistema esté vacío es $(1 - \rho)$, que es congruente con (1.7).

Por la propiedad Markoviana del tiempo de servicio, un cliente que llega y encuentra n clientes en el sistema, deberá esperar un tiempo en cola igual a la suma de n tiempos exponenciales de tasa μ, es decir, la distribución del tiempo de servicio residual es también exponencial de parámetro μ . Por tanto, condicionando en todos los posibles valores del número de clientes en el sistema, N, con distribución geométrica, (1.10), se obtiene la distribución de W, que es la distribución estacionaria del tiempo de espera en cola, $W(\tau)$, (1.4). Su función de distribución viene dada por, véase, por ejemplo, Gross y Harris (1985),

$$F_W(x) = P(W \leq x) = 1 - \rho \exp\{-\mu(1 - \rho)\,x\}, \qquad x \geq 0. \tag{1.11}$$

Obsérvese que esta distribución tiene masa positiva en el cero, concretamente, la probabilidad de que el tiempo de espera en cola sea nulo, es igual a $(1-\rho)$, que coincide con la probabilidad de que no haya clientes en el sistema.

Como se ha mencionado anteriormente, no existen muchos resultados sobre el comportamiento transitorio en sistemas de colas. Sin embargo, el sistema M/M/1 es uno de los pocos modelos para los que se conoce una expresión explícita de la distribución transitoria del tamaño del sistema, $N(\tau)$, que se obtiene resolviendo, para cada instante, τ, las ecuaciones diferenciales asociadas al proceso de nacimiento y muerte citado más arriba. La función de densidad transitoria viene dada por,

$$P(N(\tau) = n) = e^{-(\lambda+\mu)\tau}\{\rho^{\frac{n-i}{2}} I_{n-i}(2\tau\sqrt{\lambda\mu}) + \rho^{\frac{n-i-1}{2}} I_{n+i+1}(2\tau\sqrt{\lambda\mu}) + (1-\rho)\,\rho^n \sum_{j=n+i+2}^{\infty} \rho^{-j/2} I_j(2\tau\sqrt{\lambda\mu})\},$$

donde $I_n(x)$ es la función de Bessel de primer orden modificada, que viene dada por,

$$I_n(y) = \sum_{k=0}^{\infty} \frac{(y/2)^{2k+n}}{k!\,(k+n)!}, \tag{1.12}$$

véase, por ejemplo, Gross y Harris (1985).

La distribución de la longitud del periodo de ocupación se obtiene imponiendo una barrera absorbente en las ecuaciones diferenciales asociadas al proceso de nacimiento y muerte cuando el sistemas se queda vacío y estableciendo la condición inicial, $P(N(0) = 0) = 1$. De este modo, se obtiene la función de densidad del periodo de ocupación que viene dada por, véase, por ejemplo, Gross y Harris (1985),

$$f_B(x) = \frac{\exp\{-(\mu + \lambda)\,x\}}{x\rho^{1/2}} I_1\left(2x(\lambda\mu)^{1/2}\right), \qquad x \geq 0, \tag{1.13}$$

que depende de la función de Bessel, I_1, obtenida a partir de (1.12).

En el sistema M/M/1, como en todos los modelos con proceso de llegadas de Poisson, la longitud de los periodos de desocupación se distribuye según una exponencial de tasa λ.

1.3. Sistemas de colas no Markovianos.

A pesar de las buenas propiedades de los sistemas de colas Markovianos, como es natural, no es siempre posible encontrar una cadena de Markov en tiempo continuo que permita una descripción adecuada del

modelo de colas. En esta Sección, se describen, en primer lugar, los sistemas M/G/1 y GI/M/c y algunas de sus propiedades más conocidas. A continuación, una breve introducción a los métodos matriciales y a la familia de distribuciones de tipo PH, que constituyen una extensión del método de las etapas. Por último, se exponen de forma resumida algunos resultados relativos al sistema de colas general GI/G/1 y se incluyen algunos comentarios generales.

1.3.1. Los sistemas M/G/1 y GI/M/c.

Existe una clase de modelos de colas que, aunque no permiten describirse mediante una cadena de Markov en tiempo continuo, sí que pueden representarse mediante un proceso (no Markoviano) que tiene subyacente una cadena de Markov de parámetro discreto. En esta familia de modelos se incluyen los sistemas M/G/1 y GI/M/c, que reciben el nombre de modelos semi-Markovianos. En estos casos, en lugar de observar los estados de la cola en cualquier instante, τ, la idea básica consiste en describir el funcionamiento del sistema únicamente en determinados instantes y construir de este modo una cadena de Markov de parámetro discreto. Para ello, en el modelo M/G/1, se considera el número de clientes que permanecen en el sistema, $N(\tau_i)$, en los instantes consecutivos $\tau_1, \tau_2, \tau_3, ...$, inmediatamente después del abandono de cada cliente servido. Es sencillo comprobar que la cadena dada por $N_i = N(\tau_i)$, es una cadena de Markov, véase por ejemplo, Gross y Harris (1985). Igualmente, en el modelo GI/M/1, se puede considerar el número de clientes en el sistema, $N(\tau_i^*)$, que se encuentra un cliente que llega en el instante τ_i^*, (sin contarse a sí mismo). La cadena dada por $N_i^* = N(\tau_i^*)$ es también de Markov. Nótese que únicamente en los sistemas que verifican la denominada propiedad PASTA, que son aquellos cuyo proceso de llegadas es de Poisson, la distribución del número de clientes en el sistema que encuentra un cliente que llega, $N(\tau_i^*)$, coincide con la del número de clientes que se observan en un instante aleatorio, $N(\tau_i)$, véase Allen (1990) (pág. 316), donde se describe un ejemplo ilustrativo de este fenómeno. La base para el estudio del modelo GI/M/1 se generaliza de manera natural a los sistemas GI/M/c. Sin embargo, este no es el caso para el modelo M/G/1, ya que, en el sistema M/G/c, el número de llegadas en el periodo que transcurre entre las salidas de dos clientes consecutivos no depende exclusivamente del número de clientes en el instante en el que abandona el primer cliente.

Basándose en las propiedades de la cadenas de Markov de parámetro discreto, se han obtenido diferentes resultados sobre las características de los sistemas M/G/1 y GI/M/c. Algunos de ellos, que son relevantes para el contenido de esta tesis, se incluyen a continuación. Sus demostraciones y más resultados se pueden encontrar, por ejemplo, en Gross y Harris (1985).

En el modelo M/G/1, la distribución en equilibrio del número de clientes presentes verifica,

$$\Pr(N=n) = \Pr(N=0)\Pr(N_1=n) + \sum_{m=1}^{n+1} \Pr(N=m)\Pr(N_1=m-n+1), \qquad \text{para } n \geq 1, \tag{1.14}$$

con $P(N=0) = 1-\rho$, y donde N_1, es el número de llegadas durante un tiempo de servicio,

$$P(N_1=n) = \int_0^\infty \frac{(\lambda x)^n}{n!} e^{-\lambda x} f_S(x) dx,$$

donde $f_S(x)$ es la función de densidad del tiempo general de servicio y λ es la tasa del proceso Poisson de llegadas. Además, en este sistema, la distribución de equilibrio del tiempo de espera en cola, W, se conoce en términos de su transformada de Laplace-Stieljes, definida en (1.16), y viene dada por,

$$f_W^*(s) = \frac{(1-\rho)s}{s - \lambda(1 - f_S^*(s))}, \tag{1.15}$$

donde $f_S^*(s)$ es la transformada de Laplace-Stieljes de la distribución general de servicio y donde la transformada de Laplace-Stieljes de una variable aleatoria X que toma valores en $x > 0$ con función de densidad

$f(x)$ viene dada por,

$$f_X^*(s) = \int_0^\infty e^{-sx} f(x)\, dx. \tag{1.16}$$

Más generalmente, se define la transformada de Laplace de una función $f(x)$ evaluada en $x > 0$, por,

$$f^*(s) = \int_0^\infty e^{-sx} f(x)\, dx. \tag{1.17}$$

Se han obtenido también algunos resultados en términos de la transformada de Laplace de la longitud del periodo de ocupación en el sistema M/G/1 que, aunque no son expresiones cerradas, permiten obtener sus momentos. Además, se conocen los momentos de otras variables aleatorias como del número de clientes servidos durante un periodo de ocupación de este sistema. Aunque estos resultados no se pueden extender a sistemas M/G/c, existen resultados relativos a estos sistemas y a los modelos M/G/c/K, con capacidad finita. En particular, una propiedad notable en los sistemas de pérdida M/G/c/c es que el tamaño del sistema N, que coincide con el número de servidores ocupados, N_b, se distribuye como una Poisson truncada,

$$P(N_b = n) = \frac{\rho^n/n!}{\sum_{k=0}^{c} \rho^k/k!}, \qquad n = 0, ..., c. \tag{1.18}$$

En el modelo GI/M/1, la distribución estacionaria que describe el número de clientes presentes en el sistema en los instantes de llegada es una distribución geométrica que viene dada por,

$$P(N^* = n) = (1-\sigma)\sigma^j, \qquad n = 0, 1, 2, ... \tag{1.19}$$

donde σ es la única raíz en el intervalo $(0, 1)$ de la ecuación,

$$\sigma = f_A^*(\mu(1-\sigma)),$$

donde $f_A^*(s)$ es la transformada de Laplace-Stieljes, (1.16), de la distribución del tiempo entre llegadas y la distribución estacionaria del número de clientes en un instante aleatorio es,

$$P(N = n) = \begin{cases} 1-\rho, & \text{si } n = 0, \\ \rho P(N^* = n-1), & \text{si } n \geq 1. \end{cases}$$

En este sistema, la probabilidad de que un cliente tenga que esperar para ser atendido es $\sigma = 1 - P(N^* = 0)$. Usando este hecho y la distribución del tiempo de espera en la cola condicionada a que el cliente debe de esperar, se obtiene que la distribución del tiempo de espera en la cola incondicional es exponencial con un salto de tamaño $1-\sigma$ en el origen. Su función de distribución es,

$$F_W(x) = 1 - \sigma \exp\{-\mu(1-\sigma)x\}, \qquad x \geq 0.$$

Las extensiones de estos resultados para el modelo GI/M/c se incluyen en el Capítulo 4 de esta tesis.

1.3.2. Distribuciones de tipo PH y los métodos matriciales.

Los resultados que se conocen para los sistemas M/G/1 y GI/M/c son muy útiles cuando se sabe cuál es la estructura de la distribución del tiempo de servicio y entre llegadas, respectivamente. Concretamente, si sus funciones de densidad y transformadas de Laplace-Stieljes son conocidas. Sin embargo, si éste no es el caso, los resultados mencionados anteriormente no permiten obtener espresiones explícitas de las cantidades de interés. Por tanto, es importante disponer de una familia de densidades que sea lo suficientemente flexible como para aproximar cualquier distribución y que además, esté perfectamente parametrizada con el fin de calcular las medidas asociadas al sistema de colas. Éste es el caso de la clase de distribuciones de tipo PH.

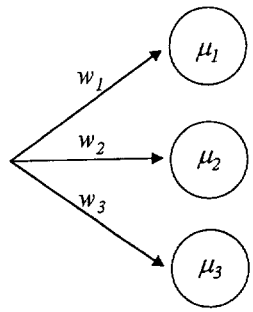


Figura 1.4: Representación de una mixtura de tres exponenciales, H_3. Cada observación puede distribuirse según una exponencial de tasa μ_1, μ_2 ó μ_3, con probabilidad w_1, w_2 y w_3, respectivamente.

La familia de distribuciones de tipo PH fue introducida por Neuts (1981) y surge como la extensión de la característica propia de la distribución Erlang, indicada en la Sección 1.2., que permite descomponer el tiempo entre llegadas o de servicio en una sucesión de fases consecutivas i.i.d. con distribución exponencial. Una distribución de tipo PH se caracteriza por corresponderse con la distribución del tiempo hasta la absorción en una cadena de Markov de parámetro continuo. Su definición y propiedades se incluyen en el Capítulo 3 de esta tesis. Esencialmente, una variable aleatoria sigue una distribución de tipo PH si se puede describir mediante un conjunto de fases exponenciales de modo que cada observación sea la suma de una sucesión de fases determinada de este conjunto. Como es natural, la distribución Erlang es de tipo PH, véase la Figura 1.3. Otros ejemplos sencillos de distribuciones PH son la exponencial y la mixtura de exponenciales. La Figura 1.4 ilustra un ejemplo de la descomposición en fases de una mixtura de tres exponenciales. Los modelos de mixtura utilizados en esta tesis para aproximar distribuciones generales del tiempo entre llegadas o de servicio en diferentes sistemas de colas son también de tipo PH.

Las distribuciones de tipo PH no se utilizan exclusivamente en los sistemas M/G/1 y G/M/c, sino que se incorporan en otros modelos de colas más generales. De hecho, esta clase de distribuciones se integra en una metodología alternativa para el estudio de los modelos de colas que fue introducida por Neuts (1981, 1989) y que está basada principalmente en la explotación de las técnicas matriciales para la obtención de expresiones explícitas de las distribuciones de interés en un sistema de colas. La idea consiste básicamente en reemplazar las distribuciones exponenciales asociadas al tiempo entre las llegadas al sistema y/o el tiempo de servicio por la clase de distribuciones de tipo PH. Por ejemplo, la cadena de Markov subyacente en el modelo de colas GI/M/1 se generaliza al modelo GI/PH/1 considerando un espacio de estados vectorial que describe, no sólo el número de clientes que encuentra un cliente que llega, sino también la fase de servicio en la que se halla el cliente que está siendo atendido. Las soluciones para este sistema tienen una estructura geométrica, similar a la que se indica en (1.19) para el modelo GI/M/1, reemplazando los escalares por matrices. Por este motivo, Neuts denomina a este tipo de soluciones geométricas matriciales (*matrix geometrics solutions*). La solución que da lugar a la distribución estacionaria asociada a los estados de la cola GI/PH/1 permite obtener también la distribución en equilibrio del tiempo de espera en cola. En Neuts (1981), se aborda también la extensión al sistema GI/PH/c. Además, se puede demostrar que el tiempo de espera en cola en el sistema GI/PH/c es también de tipo PH, véase Asmussen y Moller (2001). Por otro lado, en Ramaswami (1982), se analiza, mediante las técnicas matriciales de Neuts, la distribución del periodo de ocupación en sistemas G/PH/1, con tiempos entre llegadas no necesariamente independientes.

Haciendo uso de los métodos matriciales, en Neuts (1981, 1989), se obtienen muchos otros resultados aplicables a diferentes modelos de colas. Además, su representación matricial evita engorrosos problemas de integración numérica en los algoritmos implicados para la explicación de los modelos. Un ejemplo son los relativos a los quasi-procesos de nacimiento y muerte (*quasi-birth-death processes*). Estos procesos surgen como un caso particular muy interesante de los sistemas GI/PH/1, en el que el proceso de llegadas es

de Poisson, es decir, los modelos M/PH/1. El generador infinitesimal que describe estos procesos tiene una estructura tridiagonal por bloques que puede interpretarse como la extensión matricial de los sistemas Markovianos de tipo nacimiento y muerte, introducidos en la Sección 1.2. Además, en Neuts (1981), se demuestra que la distribución del tiempo de espera en un sistema M/PH/1 es también de tipo PH y en Neuts (1977), que la distribución de la longitud del periodo de ocupación en este sistema se puede considerar como la distribución del tiempo hasta la absorción en un proceso de Markov con un número infinito de estados, que es una versión infinita de la definición de distribución de tipo PH. En el Capítulo 3 de esta tesis, se describen detalladamente estos resultados y se incorporan en los diferentes procedimientos de estimación que se proponen.

1.3.3. Comportamiento transitorio y otros aspectos en sistemas generales.

Como se ha comentado anteriormente, en la mayoría de los modelos de colas, es difícil obtener la distribución transitoria de sus medidas de interés, por ejemplo, la distribución el número de clientes, $N(\tau)$, presentes en el sistema en cualquier instante τ. En modelos Markovianos, el problema surge de la necesidad de resolver un sistema lineal de ecuaciones diferenciales, que son las ecuaciones de Kolmogorov y que se han abordado, tradicionalmente, mediante diferentes métodos numéricos de integración, como el método de Runge Kutta o el denominado método de aleatorización (*randomization technique*), véase, por ejemplo, Grassman (1977). Las dificultades en los sistemas Markovianos se extienden a los modelos de colas semi-Markovianos, M/G/1 y GI/M/c, con las ecuaciones de Champman-Kolmogorov asociadas a las cadenas de Markov de parámetro discreto, véase Neuts (1973).

En Bertsimas y Nakazato (1992), se obtienen expresiones cerradas para las transformadas de Laplace de las distribuciones transitorias del número de clientes en el sistema, $N(\tau)$, y del tiempo de espera en cola, $W(\tau)$, de un cliente que llega en el instante, τ, en sistemas que aproximan arbitrariamente el modelo GI/G/1. Además, se obtiene la transformada de Laplace-Stieljes de la longitud del periodo de ocupación, lo que constituye una alternativa al procedimiento propuesto en Ramaswami (1982) para sistemas G/PH/1. El esquema de trabajo en Bertsimas y Nakazato (1992) está basado también en el método de las etapas, ya que la estrategia fundamental consiste en aproximar las distribuciones generales del tiempo entre las llegadas y de servicio con el modelo de mixtura denominado MGE (*Mixed Generalized Erlang*) que constituye un subconjunto muy extenso de la clase de distribuciones de tipo PH. La distribución MGE se denomina frecuentemente distribución Coxiana y su definición precisa y propiedades se incluyen en los Capítulo 2 y 3. En Bertsimas y Nakazato (1992), se aproxima el modelo GI/G/1 mediante la cola MGE/MGE/1, de modo que se obtiene un sistema Markoviano atendiendo a la definición proporcionada anteriormente. Como sucede en otros modelos de colas ya comentados, los resultados aparecen en términos de transformadas de Laplace y dada su complejidad, se hace necesario el uso de métodos de inversión numérica de transformadas para obtener expresiones explícitas. Además, como es frecuente en Teoría de Colas, se requieren técnicas para la extracción de raíces en ecuaciones complejas. Los procedimientos propuestos por Bertsimas y Nakazato (1992) se utilizan en el Capítulo 5 de esta tesis para desarrollar inferencia sobre el comportamiento transitorio y el periodo de ocupación en sistemas GI/G/1.

Para el sistema GI/G/1 sin asumir ningún modelo para las distribuciones generales, existe un resultado clave en la Teoría de Colas clásica que se conoce como la *Ecuación de Lindley* y que proporciona una ecuación integral para la distribución estacionaria del tiempo de espera en cola en instantes aleatorios. Como es natural, en la práctica, la complejidad de los procedimientos para resolver esta ecuación, que pertenece a la clase de ecuaciones denominadas de Wiener-Hopf, depende básicamente de la estructura del tiempo entre llegadas y de servicio, véase, por ejemplo, Kleinrock (1975), y se centra de nuevo en la extracción de raíces de funciones complejas.

Existen muchos otros resultados relativos a sistemas de colas generales, así como expresiones conocidas para medidas asociadas a modelos concretos diferentes a los que se han mencionado hasta ahora, tales como sistemas con diferentes disciplinas, con impaciencia, etc. Por otro lado, en la Teoría de Colas clásica, existen

numerosas técnicas diferentes y procedimientos numéricos para el análisis de modelos de colas generales, así como métodos basados en la simulación de sistemas que no se han citado en este Capítulo. Además, recientemente, se ha incrementado el análisis de estructuras más complejas, como son las redes de colas, en las que cada cliente pede requerir el servicio de más de un servidor. El volumen de trabajos en este sentido es tan numeroso que su estudio se constituye como un área de investigación diferente.

No es el objetivo de este Capítulo mencionar todos los resultados y técnicas desarrollados en la Teoría de Colas clásica. La intención en esta Sección y las anteriores ha sido reflejar fundamentalmente el material básico que ha sido utilizado en esta tesis para el desarrollo de la inferencia Bayesiana en los sistemas de colas generales que se consideran.

1.4. Análisis estadístico para sistemas de colas.

En las Secciones anteriores se ha presentado una visión global de algunos de los resultados clásicos de la Teoría de Colas. Tal como se ha indicado, para la derivación de estos resultados se asume que tanto el modelo de colas como sus parámetros son conocidos. Como ya se ha comentado, en la práctica, no se dispone de esta información y se hace necesario el uso de métodos estadísticos que permitan la estimación de las cantidades de interés en un sistema de colas a partir de la observación del mismo.

La inferencia para modelos de colas no es una tarea sencilla teniendo en cuenta que, generalmente, se requiere la incorporación de incertidumbre sobre los parámetros que determinan un modelo cuyas características, conocidos los parámetros, no son generalmente fáciles de calcular. Para abordar el análisis estadístico de un sistema es necesario, por un lado, diseñar un experimento adecuado para la observación del mismo. Por otro lado, es importante la selección de un modelo de colas que permita describir el sistema observado apropiadamente. Como es natural, si el modelo escogido es demasiado restrictivo, puede suceder que no explique bien el funcionamiento del sistema y las estimaciones de las cantidades de interés no se ajusten a la realidad. Además, antes de abordar la inferencia y predicción en el sistema, es importante verificar si se cumplen las hipótesis de partida que se hayan establecido en el modelo de colas, tales como la independencia entre las llegadas y servicios.

1.4.1. Algunos procedimientos clásicos de estimación en modelos de colas

Como se comentó anteriormente, en contraste con el volumen de trabajos destinados al análisis probabilístico de los modelos de colas, el estudio estadístico de los mismos ha recibido muy poca atención. Se puede encontrar una revisión general sobre el tema en Bhat et al. (1997).

La mayoría de los análisis basados en métodos de inferencia clásica se centran en la estimación puntual de los parámetros asociados al proceso de llegadas y de servicio, que se ha desarrollado, fundamentalmente, a través de estimadores máximo verosímiles para modelos de colas Markovianos, véase, por ejemplo, Clarke (1957) y Wolff (1965). También, existen algunos trabajos en los que se consideran otros procedimientos de estimación, como el método de los momentos, así como algunos modelos de colas no Markovianos, véase Gross y Harris (1985) y las referencias que allí se indican. Con métodos clásicos, no se ha desarrollado únicamente estimación puntual, por ejemplo, se han obtenido también intervalos de confianza para la tasa de llegadas y de servicio, así como para la intensidad de tráfico en modelos Markovianos, véase, por ejemplo, el trabajo recopilatorio de Cox (1966), donde se proponen también contrastes de hipótesis para los parámetros mencionados. Por último, sin asumir ningún modelo de colas, se han derivado estimadores de algunas cantidades de interés a partir de su observación directa, como, por ejemplo, el número medio de clientes en el sistema o el tiempo medio de espera en cola, obtenidos en Halfin (1982) y Blomqvist (1967), respectivamente.

Por otro lado, otra metodología más reciente para la estimación en modelos de colas es la simulación estocástica. Estas técnicas se utilizan generalmente cuando no es posible encontrar modelos analíticos que

describan los sistemas de colas que se pretenden analizar. La estrategia básicamente es la siguiente. Se generan valores aleatorios de las distribuciones del tiempo de servicio o de llegadas (a menudo obtenidas a partir de estimaciones empíricas de la función de distribución de los datos observados). Con los valores generados, se simula la dinámica del sistema almacenando la información necesaria para, posteriormente, estimar las cantidades de interés. Gross y Harris (1985) advierte del cuidado especial que es necesario en los procedimientos de validez estadística de estas estimaciones. La mayoría de los trabajos se centran en el cálculo de los valores medios de estas medidas, como, por ejemplo, el número medio de clientes en el sistema. Algunas referencias en esta línea de trabajo se pueden encontrar en Law y Kelton (1991), Banks et al. (1996) y Fishman (2001).

1.4.2. Enfoque Bayesiano para el análisis sistemas de colas.

Como ya se ha indicado, en esta tesis, se ha optado por la metodología Bayesiana para el análisis estadístico de los sistemas de colas. En el siguiente apartado, se exponen algunas de las cualidades que ofrece esta perspectiva en el contexto de colas. A continuación, en el apartado 1.4.2.2, se incluyen una revisión de la literatura Bayesiana para el análisis de colas. Por ultimo, en el apartado 1.4.2.3, se describe brevemente cómo desarrollar inferencia y predicción Bayesiana para el sistema M/M/1.

1.4.2.1. Algunas cualidades de la perspectiva Bayesiana en colas.

En Armero y Bayarri (1999), se describen detalladamente las ventajas e inconvenientes que ofrece la perspectiva Bayesiana para el análisis de los sistemas de colas. Las razones que se exponen en esta referencia para optar por esta metodología se exponen de forma esquemática a continuación.

- El principio de verosimilitud.

  La inferencia Bayesiana, como la estimación máximo verosímil y a diferencia de los contrastes o intervalos de confianza clásicos, verifica el principio de verosimilitud, que establece que toda la información que un experimento pueda proveer está resumida en la función de verosimilitud. Este principio conduce a la selección de experimentos sencillos para la observación del sistema de colas, como el que se indica en la Sección 1.4.2.3, para el modelo M/M/1, que se ha utilizado también para en estimaciones máximo verosímiles, véase, por ejemplo, Thiruvaiyaru y Basawa (1992).

- Probabilidades de interés.

  Puesto que, desde el punto de vista Bayesiano, los parámetros son variables aleatorias, muchas de las probabilidades que interesan en un modelo de colas se pueden calcular fácilmente a partir de las distribuciones a posteriori de los parámetros del sistema. Una de las más importantes es la probabilidad a posteriori de que se verifique la condición de ergodicidad, dada en (1.6).

- Restricciones en el espacio paramétrico.

  Con un enfoque Bayesiano, la inclusión de restricciones en el espacio de parámetros es sencilla ya que se pueden incorporar directamente en las distribuciones a priori de los mismos. En particular, en sistemas de colas, cuando se pretende analizar el modelo en equilibrio es necesario imponer la condición de ergodicidad, $\rho < c$, dada en (1.6), para que existan las distribuciones estacionarias, lo cual se puede incluir directamente en un análisis Bayesiano.

- Precisión de las estimaciones.

  La perspectiva Bayesiana, en general, permite calcular de forma natural la precisión de las estimaciones a partir de la varianza de la distribución a posteriori de los parámetros, entre otras medidas. Este fenómeno se aplica en sistemas de colas, permitiendo calcular, por ejemplo, no sólo la varianza a

posteriori de la intensidad de tráfico, ρ, (1.5), sino también el valor de este varianza condicionado a que el sistema sea estable.

- Predicción.

  Este es posiblemente uno de los puntos fuertes de esta relación. Como se comentó en la Sección 1.1.2, el interés en un modelo de colas se centra fundamentalmente en las medidas asociadas al mismo, tales como el número de clientes en el sistema en equilibrio, N. Un procedimiento frecuente en la estimación puntual clásica es su incorporación directa en las cantidades de interés, por ejemplo, si se tiene un estimación $\hat{\theta}$ de los parámetros que determinan el sistema de colas, estimar la distribución de N mediante $P(N = n \mid \hat{\theta})$. Schruben y Kulkarni (1982) advierten de los problemas que pueden surgir si no se tiene en cuenta la varianza de las estimaciones de los parámetros del sistema cuando se estiman estas medidas de interés. El enfoque Bayesiano permite la incorporación directa de la incertidumbre en las estimaciones a través de las distribuciones predictivas calculando, por ejemplo,

  $$P(N = n \mid datos) = \int_{\Theta} P(N = n \mid \theta) f(\theta \mid datos) \, d\theta,$$

  donde Θ representa el espacio paramétrico de θ.

- Análisis transitorio.

  Desde el punto de vista conceptual, la estimación del comportamiento transitorio del sistema no ofrece dificultades en los procedimientos Bayesianos. Sin embargo, puesto que, en general, este análisis es complicado si los parámetros del sistema son conocidos, la dificultad se incrementa aún más cuando se incorporan en procedimientos de estimación, sea cual sea el enfoque considerado. Nótese que, en este caso, no es necesario asumir equilibrio en el sistema. Más comentarios en este sentido se incluyen en el Capítulo 5.

- Diseño.

  En el diseño de sistemas de colas, la opinión de los especialistas es fundamental y, como es sabido, los métodos Bayesianos permiten su incorporación natural en los procedimientos de estimación y decisión. Además, el análisis Bayesiano permite la formulación de funciones o estructuras de costes (o pérdida) para la toma de decisiones que incorporen la incertidumbre de las estimaciones. Estos procedimientos se describen detalladamente en el Capítulo 4.

Evidentemente, estos comentarios se extienden a los modelos de colas más generales considerados a lo largo de esta tesis. Además, como se mostrará más adelante, la metodología Bayesiana permite, haciendo uso de los métodos MCMC, estimar las distribuciones generales y desconocidas del tiempo entre llegadas y/o de servicio en el sistema, así como incorporar esta incertidumbre en las distribuciones predictivas de las cantidades de interés, como el tamaño del sistema, en sistemas generales.

En Armero y Bayarri (1999), se incluyen también algunas dificultades y cuestiones sin resolver que aparecen en la inferencia Bayesiana para sistemas de colas, que se centran, fundamentalmente, en el problema de la selección de una distribución a priori adecuada para los parámetros del modelo. Concretamente, se advierte de las complicaciones que pueden surgir cuando se seleccionan, sin otro motivo que la simplicidad, distribuciones a priori no informativas y/o conjugadas, así como de la necesidad de plantear distribuciones apropiadas que reflejen el conocimiento a priori que se tiene sobre el equilibrio del sistema. En Armero y Bayarri (1999), se describe también el problema de la no existencia de momentos en la mayoría de las distribuciones predictivas de las cantidades de interés de un sistema de colas en equilibrio. Este fenómeno surge cuando, entre otras condiciones, se asume una distribución a priori para la intensidad de tráfico, ρ, con masa positiva en $\rho = c$, ya que la esperanza a posteriori de estas distribuciones es una media ponderada de cantidades que tienden a infinito cuando ρ tiende a c. Como se mostrará más adelante, esta dificultad aparece también en modelos de colas más generales. En Armero y Bayarri (1994b) se proponen distribuciones

alternativas a priori que eviten este problema, aunque, en general, no son fáciles de extender a otros modelos de colas más generales. Hay que puntualizar que éste no es un problema exclusivo del enfoque Bayesiano, como se muestra, por ejemplo, en Schruben y Kulkarni (1982).

1.4.2.2. Revisión de la literatura Bayesiana en sistemas de colas.

Actualmente, la inferencia Bayesiana en sistemas de colas es un área de trabajo bastante desarrollada. Sin embargo, en la mayoría de los casos, el análisis se ha centrado en modelos de colas Markovianos. En esta Sección, se incluye una recapitulación de los análisis Bayesianos más relevantes que se han desarrollado hasta el momento.

El análisis de modelos de colas desde una perspectiva Bayesiana parece remontarse a principios de los años setenta. Los primeros trabajos abordan el problema de la estimación puntual de la tasa de llegadas y de servicio en sistemas de colas Markovianos mediante estimadores Bayes, véase Muddapur (1972) y Reynolds (1973). En la misma época, Bagchi y Cunningham (1972) proponen algunos procedimientos Bayesianos para el diseño óptimo de sistemas de colas con un único servidor con el fin de encontrar la mejor tasa de servicio y capacidad del sistema según unos costes preestablecidos. Algunas de sus ideas se utilizan en esta tesis, en el Capítulo 4, para la optimización sobre el número de servidores.

A mediados de los años ochenta, se produce un resurgimiento del interés en el tratamiento Bayesiano de los sistemas de colas. En Armero (1985), se obtiene la distribución a posteriori de la intensidad de tráfico y las distribuciones predictivas del número de clientes presentes en el sistema y del tiempo de espera en cola para el modelo M/M/1 en equilibrio. Por otro lado, McGrath et al. (1987) y McGrath y Singpurwalla (1987), en dos artículos consecutivos, desarrollan un trabajo extenso para establecer conceptos básicos en el análisis Bayesiano de la incertidumbre en sistemas de colas y desarrollar inferencia en el modelo M/M/1/K, con $K \leq \infty$, mediante la información de Shannon y con diferentes experimentos para la observación del sistema, uno de ellos es el considerado en esta tesis que se describe en la Sección 2.1 del Cap 2.

Desde principios de los años noventa hasta la época actual, el número de trabajos en los que se utilizan las técnicas Bayesianas para sistemas de colas ha crecido considerablemente. En Thiruvaiyaru y Basawa (1992), se desarrolla estimación Bayesiana empírica (*empirical Bayes estimation*), basada en la estimación puntual de los hiperparámetros, sobre diferentes modelos de colas Markovianos y redes de colas Jackson abiertas. En Armero y Bayarri (1994a), se completa el análisis desarrollado en Armero (1985), para el sistema M/M/1, obteniendo, entre otros resultados, la distribución predictiva de la longitud del periodo de ocupación. Además, se expone el problema de la no existencia de momentos en la mayoría de las distribuciones predictivas de las características del sistema cuando se utilizan distribuciones a priori conjugadas. Este fenómeno motiva, en Armero y Bayarri (1994b), la selección de una familia de distribuciones a priori apropiada para la predicción en el sistema M/M/1. Por otro lado, Armero (1994) examina la condición de ergodicidad como un problema de decisión en modelos de colas Markovianos. Más adelante, el estudio estadístico del sistema M/M/1 se generaliza, en diversos trabajos, a otros sistemas de colas Markovianos. Por ejemplo, en Armero y Bayarri (1996), se desarrolla inferencia y predicción Bayesiana para el modelo M/M/c y se estudian diferentes criterios para la decisión de un número apropiado de servidores en este sistema. En Armero y Bayarri (1997), se considera el modelo de colas M/M/∞ utilizando varios experimentos para su observación y se analiza, además, su comportamiento transitorio. Por otro lado, en Rodrigues y Galvão (1998), se obtiene la distribución a posteriori de la intensidad de tráfico en un sistema M/M/1 a partir de una función de verosimilitud especial para ρ, los resultados se comparan con los de Armero y Bayarri (1994b).

Además de esta colección de trabajos centrados, fundamentalmente, en los sistemas Markovianos M/M, se han desarrollado diferentes aportaciones Bayesianas para otros modelos de colas más generales. Concretamente, relajando el requerimiento de que la distribución del tiempo de servicio sea exponencial, se han aproximado algunas de las cantidades de interés de varios sistemas M/G/1. Por ejemplo, en Bhattacharya y Singh (1994), se estima la intensidad de tráfico del sistema M/Er/1. En Ríos et al. (1998), se aproximan

además las distribuciones predictivas del tamaño del sistema y del tiempo de espera en cola en los modelos M/Er/1 y M/H_k/1. Para ello, se hace uso de los métodos MCMC con dimensión paramétrica fija o variable, según el caso. En Wiper et al. (2001), se generalizan algunos de estos procedimientos a sistemas de colas en los que el tiempo de servicio se modela con una mixtura de distribuciones Gamma. Puesto que este modelo de mixtura no es asequible para estimar algunas de las medidas relevantes del sistema, en Ausín et al. (2004), se considera una mixtura de distribuciones Erlang para la distribución del tiempo de servicio lo cual permite hacer uso de las buenas propiedades de los sistemas M/PH/1 para predecir las distribuciones de las medidas de interés en equilibrio. Los procedimientos descritos en esta referencia están contenidos en esta tesis, en los Capítulos 2 y 3. En otra línea de trabajo, Butler y Huzurbazar (2000) analizan la distribución predictiva de los tiempos de espera en el sistema M/G/1 considerando una distribución inversa Gaussiana para el tiempo de servicio.

Existen también varias contribuciones en las que se extiende la inferencia a sistemas de colas con procesos de llegadas más complicados que el de Poisson. Por ejemplo, en Wiper (1998), se estiman varias cantidades de interés del sistema de colas Er/M/c y se establecen, además, unas condiciones generales a priori que conducen a la no existencia de momentos de la mayoría de las distribuciones predictivas de las medidas estimadas. Por otro lado, en Armero y Conesa (1998), se desarrolla inferencia y predicción Bayesiana para el sistema de colas M^k/M/1, en el que las llegadas se producen en grupos de tamaño fijo, así como para el sistema M/Er/1 haciendo uso de la dualidad entre ambos modelos. Para abordar la inferencia en estos sistemas se utilizan técnicas de inversión numérica de transformadas de Laplace. Este trabajo se generaliza, en Armero y Conesa (2000), a modelos de colas Markovianos en los que las llegadas se producen en grupos de tamaño variable.

Las extensiones en el análisis Bayesiano de sistemas de colas no se ha reducido exclusivamente a la generalización del proceso de llegadas o de servicio, sino también a otros aspectos como, por ejemplo, la relación de dependencia entre ambos procesos, véase Ruggeri et al. (1996), donde se supone que la tasa de servicio y de llegadas son dependientes a priori. Por otro lado, una extensión natural en la Teoría de Colas clásica es el estudio de las redes de colas el cual ha adquirido, en los últimos años, una importancia considerable por la diversidad de sus aplicaciones. Actualmente, existen varios estudios dedicados al análisis Bayesiano de estas estructuras, véanse, por ejemplo, Armero y Bayarri (1999), Tebaldi y West (1997) y Ganesh et al. (1998).

Por último, la metodología Bayesiana se ha aplicado para resolver problemas que se plantean en diferentes situaciones reales que permiten modelarse mediante sistemas de colas. Por ejemplo, en Ausín et al. (2003), se analizan diversas cuestiones relacionadas con la gestión hospitalaria mediante modelos de colas M/G/c/c. El contenido de esta referencia se incluye en esta tesis, en los Capítulos 2 y 4. Otro ejemplo se puede encontrar en Armero y Conesa (2003), donde se examina la gestión de existencias en sistemas de producción basándose en modelos de colas con procesos múltiples de llegadas en grupos.

1.4.2.3. Inferencia Bayesiana para el sistema M/M/1.

La metodología considerada en Armero y Bayarri (1994a) para el análisis Bayesiano del sistema M/M/1 constituye uno de los pilares fundamentales para el trabajo desarrollado en esta tesis. Este es el motivo por el que, en esta Subsección, se resume, muy brevemente, el procedimiento propuesto para la inferencia y predicción Bayesiana en el mencionado sistema, que es el modelo de colas más sencillo.

El experimento para la observación del sistema, que se considera también en esta tesis y se detalla en la Sección 2.1, consiste básicamente en la observación de n_s tiempos de servicio, $\mathbf{s} = \{s_1, ..., s_{n_s}\}$, y n_a tiempos entre llegadas, $\mathbf{t} = \{t_1, ..., t_{n_a}\}$. Puesto que, como se describe en la Sección 1.2, los datos, $\mathbf{s}$ y $\mathbf{t}$, son dos conjuntos independientes de observaciones i.i.d. exponencialmente distribuidas con tasas λ y μ, respectivamente, no es necesario que se observen simultáneamente y la función de verosimilitud viene dada

por,

$$L(\lambda,\mu \mid \mathbf{t},\mathbf{s}) = \lambda^{n_a}\exp\{-\lambda\textstyle\sum_{i=1}^{n_a} t_i\}\,\mu^{n_s}\exp\{-\mu\sum_{i=1}^{n_s} s_i\}.$$

En Armero y Bayarri (1994a), se estudian los aspectos de interés prediciendo bajo distribuciones a priori naturales conjugadas. Concretamente, se suponen distribuciones Gamma a priori, $\lambda \sim G(a_\lambda^0, b_\lambda^0)$ y $\mu \sim G(a_\mu^0, b_\mu^0)$, para las tasas de llegadas y de servicio, respectivamente. Es fácil comprobar que, con esta estructura, las distribuciones a posteriori de λ y μ son independientes y vienen dadas por,

$$\lambda \mid \mathbf{t} \sim G\left(a_\lambda^0 + n_a, b_\lambda^0 + \textstyle\sum_{i=1}^{n_a} t_i\right), \tag{1.20}$$

y,

$$\mu \mid \mathbf{s} \sim G\left(a_\mu^0 + n_s, b_\mu^0 + \textstyle\sum_{i=1}^{n_s} s_i\right). \tag{1.21}$$

Además, a partir de (1.20) y de (1.21), se demuestra que la distribución a posteriori de la intensidad de tráfico, ρ, dada en (1.9), se expresa en términos de una distribución F de Fisher,

$$\frac{\rho}{R} \mid \mathbf{t},\mathbf{s} \sim F(2a_\lambda, 2a_\mu), \tag{1.22}$$

donde,

$$R = \frac{a_\lambda}{b_\lambda}\frac{a_\mu}{b_\mu} = \frac{E[\lambda \mid \mathbf{t}]}{E[\mu \mid \mathbf{s}]},$$

y donde $a_\lambda = \left(a_\lambda^0 + n_a\right)$, $a_\mu = \left(a_\mu^0 + n_s\right)$, $b_\lambda = \left(b_\lambda^0 + \sum_{i=1}^{n_a} t_i\right)$ y $b_\mu = \left(b_\mu + \sum_{i=1}^{n_s} s_i\right)$.

La distribución dada en (1.22) permitirá estudiar el comportamiento del sistema de colas y estimar algunos de sus rasgos interesantes. Por ejemplo, en Armero y Bayarri (1994a), se obtiene la probabilidad a posteriori de que el sistema sea estable,

$$P(\rho < 1 \mid \mathbf{t},\mathbf{s}) = \int_0^1 f(\rho \mid \mathbf{t},\mathbf{s})\,d\rho \tag{1.23}$$

$$= \frac{\Gamma(a_\mu + a_\lambda)\,(b_\lambda/b_\mu)^{a_\mu}}{\Gamma(a_\lambda)\Gamma(a_\mu)} F(a_\mu + a_\lambda, a_\lambda; 1 + a_\lambda; -b_\lambda/b_\mu), \tag{1.24}$$

donde F es la función hipergeométrica,

$$F(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)}\int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}dt, \qquad c > b > 0.$$

El cálculo de la probabilidad a posteriori de equilibrio en el sistema M/M/1, (1.23), se generalizará en esta tesis a otros sistemas de colas más generales. Mediante procedimientos que se describirán más adelante, se obtendrán aproximaciones Monte Carlo de la probabilidad de que el sistema sea estable generando valores de la distribución a posteriori de ρ.

Para analizar el comportamiento del sistema en equilibrio es necesario asumir que se verifica la condición de ergodicidad, $\rho < 1$, véase (1.6), ya que en caso contrario no existe la distribución estacionaria asociada a los estados del sistema cuando se conocen sus parámetros. Por este motivo, en Armero y Bayarri (1994a), se obtiene la distribución a posteriori de ρ asumiendo equilibrio, $f(\rho \mid \mathbf{t},\mathbf{s},\rho < 1)$, a partir de (1.22) y de (1.23), lo cual va a permitir calcular las distribuciones predictivas de las medidas más importantes del sistema en el estado estacionario. Concretamente, en Armero y Bayarri (1994a), se consiguen expresiones explícitas para las distribuciones predictivas del número de clientes presentes en la cola de espera y en el sistema, así como del tiempo de espera en cola y en el sistema. También, se obtienen las distribuciones predictivas de la longitud del periodo de ocupación y de desocupación del sistema. Por brevedad, se expone únicamente la distribución predictiva exacta del tamaño del sistema que viene dada por,

$$P(N = n \mid \mathbf{t},\mathbf{s},\rho < 1) = \int_0^1 P(N = n \mid \rho)\,f(\rho \mid \mathbf{t},\mathbf{s},\rho < 1)\,d\rho$$

$$= \frac{a_\lambda F(a_\mu + a_\lambda, n + a_\lambda, n + a_\lambda + 2; -b_\lambda/b_\mu)}{(n + a_\lambda + 1)(n + a_\lambda)F(a_\mu + a_\lambda, a_\lambda; 1 + a_\lambda; -b_\lambda/b_\mu)},$$

donde $P(N = n \mid \rho)$ viene dado en (1.10). Con un procedimiento similar, en Armero y Bayarri (1994a), se obtienen las restantes distribuciones predictivas. Utilizando esta formulación como base, en esta tesis, se obtendrán las distribuciones predictivas de las cantidades de interés en diferentes sistemas de colas generales. Para ello, se utilizarán algunos de las técnicas conocidas de la Teoría de Colas clásica, que se han mencionado en la Sección 1.3, para calcular las probabilidades asociadas al sistema cuando se conocen sus parámetros. Estos procedimientos se combinarán con diferentes métodos que permitirán obtener muestras de la probabilidad condicional, $f(\rho \mid \mathbf{t}, \mathbf{s}, \rho < 1)$, y obtener aproximaciones Monte Carlo de las medidas consideradas.

Por último, es importante señalar que, en Armero y Bayarri (1994a), se deriva que, bajo las condiciones a priori expuestas anteriormente, no existen los momentos de las distribuciones predictivas mencionadas, con la excepción de la longitud de los tiempos de desocupación cuyos momentos son finitos hasta el orden a_λ. Como se verá a lo largo de esta tesis, este resultado se repite en sistemas de colas con distribuciones más generales que la exponencial para el tiempo entre las llegadas y de servicio en el sistema.

Capítulo 2

Estimación Bayesiana de densidades: Inferencia para las distribuciones del proceso de llegadas y de servicio.

En este Capítulo, se aborda el primer paso para la inferencia en sistemas de colas que es el estudio del proceso de llegadas y de servicio. Estos procesos caracterizan el tipo de modelo de colas a estudiar e influyen directamente sobre el comportamiento probabilístico de las cantidades que interesan específicamente en un modelo de colas entre las que se incluyen los tiempos de espera, el número de clientes en el sistema y la longitud de los periodos de ocupación. Aunque todas ellas son magnitudes observables, no es habitual, desde una perspectiva Bayesiana, observar directamente estas medidas sino que es más eficaz tomar datos procedentes del proceso de llegadas y de servicio, caracterizar el modelo de colas y aplicar los resultados de la Teoría de Colas clásica en los que se derivan el comportamiento aleatorio de estas cantidades de interés. En Capítulos posteriores, los resultados de inferencia sobre estos procesos se incorporan para abordar los problemas de predicción de las medidas mencionadas. Puesto que se asume que las llegadas y los servicios constituyen sucesiones independientes de variables aleatorias i.i.d., este Capítulo se destina a la estimación Bayesiana de densidades de variables aleatorias continuas y positivas, con aplicaciones obvias e interesantes fuera del contexto de la tesis.

El objetivo en este Capítulo es hacer inferencia sobre el proceso de llegadas y de servicios en un sistema de colas considerando que ambos son procesos de renovación, independientes entre sí y regidos, respectivamente, por una variable aleatoria con distribución general y desconocida definida en la semirecta real positiva. Para aproximar esta variable, en algunas ocasiones, se escogen modelos de distribución tradicionales en sistemas de colas, como son la distribución exponencial, Erlang o las mixturas de exponenciales, véase, por ejemplo, Armero y Bayarri (1994a), Wiper (1998) y Ríos et al. (1998), respectivamente. Sin embargo, no es siempre evidente que estas distribuciones tan elementales aproximen bien todos los aspectos de las observaciones reales. En otras ocasiones, se escogen distribuciones difíciles de incorporar en los cálculos posteriores de las medidas implicadas en modelos de colas. Por ejemplo, los tiempos de servicio de algunos sistemas relacionados con Internet, aparecen frecuentemente modelados con colas muy pesadas, para lo cual se utilizan habitualmente la distribución Lognormal, Weibull o Pareto. Sin embargo, no es sencillo, en general, incluir estas distribuciones estimadas en un sistema de colas debido a que los cálculos asociados a sistemas con estos modelos de distribución para el proceso de servicio o de llegadas son muy laboriosos, véase, por ejemplo, Gross et al. (2002), Feldmann y Whitt (1998) y Harris (1968). Una alternativa posible sería no suponer ningún modelo paramétrico para la distribución del servicio o de llegadas y desarrollar un procedimiento similar al que se propone en Coolen y Coolen-Schrijner (2003). Sin embargo, bajo este enfoque

no paramétrico tampoco se asume un modelo estocástico para el sistema y por tanto, no se pueden integrar directamente los resultados clásicos de la Teoría de Colas para hacer predicciones, por ejemplo, sobre la longitud del periodo de ocupación. En concreto, en Coolen y Coolen-Schrijner (2003), se estiman únicamente las probabilidades predictivas del tiempo de espera asumiendo unas hipótesis mínimas sobre la distribución de servicio.

Se pretende, por tanto, seleccionar un modelo de distribución asequible operacionalmente en un contexto de colas y que, a su vez, sea lo suficientemente flexible para captar toda la variabilidad que ofrezca el proceso de llegadas y el de servicio. Con esta finalidad, la propuesta de este trabajo es utilizar modelos semiparamétricos basados en mixturas de distribuciones. Concretamente, se presentan dos modelos de mixturas que permiten aproximar arbitrariamente cualquier variable aleatoria positiva y que además gozan de muy buenas propiedades para la predicción de las medidas interesantes de los sistemas de colas. Estos modelos son la familia de mixturas de distribuciones Erlang (HEr) y el modelo de distribución Coxiana (MGE), que se describirán detalladamente en la Sección siguiente.

En este Capítulo, se va a suponer que el número de términos que intervienen en las mixturas de distribuciones citadas es desconocido. Es importante puntualizar que, en este contexto, el número de componentes de la mixtura no tiene interés en sí mismo, sino que se permite que su valor varíe con la finalidad de aumentar la flexibilidad de los modelos considerados y, en general, las componentes de la mixtura no van a tener una interpretación física. Concretamente, el problema que se plantea no es la selección de los mejores valores para los parámetros del modelo de mixtura considerado, sino la construcción de una combinación razonable de los resultados obtenidos con diferentes modelos. El enfoque Bayesiano ofrece una metodología muy natural en esta situación, ya que permite obtener la estimación de una densidad determinada como el resultado de una ponderación adecuada de diferentes estimaciones en distintas dimensiones.

Globalmente, se proponen dos líneas de actuación para desarrollar inferencia Bayesiana. La primera propuesta es un procedimiento basado en los métodos de salto reversible, introducidos por Green (1995) y adaptados posteriormente para mixturas de normales, véase Richardson y Green (1997). Esta técnica se considera también en Ríos et al. (1998) y Gruet et al. (1999) para mixturas de exponenciales y Wiper et al. (2001) para mixturas de distribuciones gamma, y fuera del contexto de mixturas, Robert et al. (2000) para modelos de cadenas de Markov ocultas. La segunda propuesta desarrolla otro procedimiento para la estimación Bayesiana de los modelos de mixtura mencionados y se basa en un trabajo reciente de Stephens (2000a). La metodología de Stephens (2000a) constituye una alternativa a los métodos de salto reversible y se ha desarrollado en un contexto de mixturas de normales. Se fundamenta en la construcción de procesos de nacimiento y muerte en tiempo continuo que tienen como distribución estacionaria la distribución a posteriori de interés en cada caso.

Este Capítulo está organizado en seis Secciones. En la Sección 2.1, se explica brevemente el experimento tradicionalmente considerado para la observación de los datos procedentes del proceso de llegadas y de servicio. En la Sección 2.2, se describen detalladamente los dos modelos de mixtura considerados, la distribución HEr y el modelo MGE, y se especifican algunas de sus propiedades más atractivas. Se ilustra, también, cómo la distribución HEr puede ser considerada como un caso particular de la distribución MGE, lo cual se comprobará formalmente en el Capítulo 3. A pesar de esta propiedad, la parametrización es distinta lo cual implica que existen diferencias importantes de cara a la inferencia. La Secciones 2.3 y 2.4 están destinadas a la descripción de los métodos propuestos para la estimación Bayesiana de densidades basados en los dos modelos de mixtura seleccionados. En particular, en la Sección 2.3, se describe un método MCMC para cada modelo de distribución considerado, y se asume, inicialmente, que el número de componentes en la mixtura es conocido. En la Sección 2.4, se incorporan los métodos de la Sección 2.3 para construir algoritmos que permitan la inferencia cuando se desconoce la dimensión de los parámetros. Se desarrolla un procedimiento para ajustar una distribución HEr y otro distinto para la distribución MGE, ambos basados en los algoritmos de salto reversible. Del mismo modo, se exponen otros dos algoritmos basados en las técnicas para tiempo continuo propuestas por Stephens (2000a). En la Sección 2.5, se incluye una extensa ilustración de los métodos descritos en las Secciones anteriores con cinco conjuntos de datos simulados, procedentes de las dos

distribuciones consideradas y de otras muy dispares, así como un conjunto de datos reales sobre estancias de pacientes en un hospital geriátrico de Londres. Además, se establece en la Sección 2.5 una comparación numérica de los métodos de las Secciones anteriores exponiendo sus ventajas e inconvenientes. Se comparan los resultados de las estimaciones y se analiza la velocidad de convergencia y la capacidad para explorar el espacio paramétrico. También se desarrolla un análisis de sensibilidad respecto a la elección de la distribución a priori. Por último, se concluye, en la Sección 2.6, con algunos comentarios y extensiones.

2.1. Experimento para la observación del sistema.

Un experimento muy habitual para obtener información procedente de un sistema de colas está basado en un procedimiento muy sencillo y que, a su vez, ofrece información completa sobre el sistema. Consiste en observar n_s tiempos de servicio, $\mathbf{s} = \{s_1, ..., s_{n_s}\}$, y n_a tiempos entre llegadas, $\mathbf{t} = \{t_1, ..., t_{n_a}\}$. No es necesario que estos procesos se observen simultáneamente. Como se considera que las llegadas y los servicios son sucesiones independientes de variables aleatorias i.i.d., la función de verosimilitud es,

$$L(\theta_\lambda, \theta_\mu \mid \mathbf{t}, \mathbf{s}) \propto L(\theta_\lambda \mid \mathbf{t})\, L(\theta_\mu \mid \mathbf{s}),$$

que consta de dos partes, una correspondiente a los parámetros del proceso de llegadas, θ_λ, y otra que corresponde a los parámetros del servicio, θ_μ. Por tanto, si se asumen distribuciones a priori independientes para los parámetros de llegadas y servicios, se obtienen distribuciones a posteriori también independientes para llegadas y servicios. Este es el experimento que se ha considerado tradicionalmente en el contexto Bayesiano para la inferencia en sistemas de colas, véase por ejemplo, Thiruvaiyaru y Basawa (1992), Ríos et al. (1998) y Armero y Bayarri (1994a). Sin embargo, se pueden usar otros procedimientos que darían lugar a formas diferentes de la verosimilitud, como los que se proponen en Lehoczky (1990). Por ejemplo, se podría observar el proceso de llegadas y de servicios durante un periodo de tiempo, $[0, T]$, fijado previamente. Si se consideran procesos de renovación más generales que el de Poisson para las llegadas o servicios, esta alternativa supondría una complejidad adicional originada por la existencia de observaciones incompletas que implicarían la incorporación de variables censuradas.

Considerando el experimento habitual y las hipótesis de independencia asumidas, se pueden estudiar por separado las distribuciones del tiempo de servicio y del tiempo entre llegadas. Por esta razón, el punto de partida en este Capítulo es una muestra de datos procedentes de un proceso de renovación que se pudiera asociar bien al proceso de llegadas, o bien al proceso de servicio, a partir de la cual se desarrollan distintos procedimientos para la inferencia sobre los parámetros y la estimación de la densidad correspondiente. Además, como se ha comentado anteriormente, se asume que la dimensión del espacio paramétrico considerado es desconocida. Como consecuencia de este comentario, se subraya la generalidad de este Capítulo que puede ser tratado fuera del contexto de la Teoría de Colas y dentro del de estimación de densidades.

2.2. Distribuciones para el proceso de servicio y de llegadas.

En esta Sección, se describen los dos modelos de distribución que se proponen para ajustar un conjunto de datos $\mathbf{x} = \{x_1, ..., x_n\}$ que, como ya se ha mencionado, pueden ser una muestra de tiempos de servicio, $\mathbf{s}$, o una muestra de tiempos entre llegadas al sistema, $\mathbf{t}$. Consecuentemente, los modelos de mixturas que se exponen son mixturas de distribuciones continuas y definidas en la semirecta real positiva.

2.2.1. Mixturas de distribuciones Erlang (HEr).

El modelo de distribución semiparamétrico basado en una mixtura de distribuciones Erlang se define del siguiente modo. Sea X una variable aleatoria distribuida según una mixtura de k distribuciones Erlang con

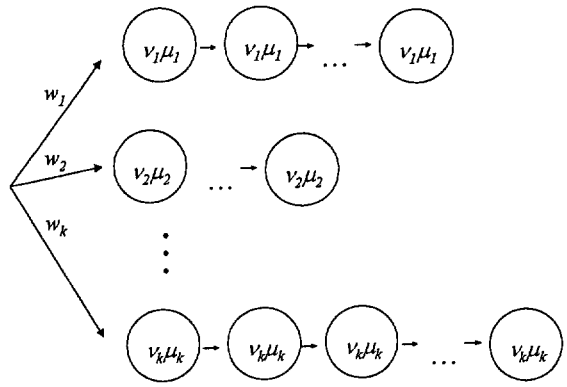


Figura 2.1: Representación de las fases exponenciales de la distribución HEr.

parámetros $\mathbf{w} = (w_1, ..., w_k)$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)$ y $\boldsymbol{\nu} = (\nu_1, ..., \nu_k)$. Su correspondiente función de densidad viene dada por,

$$f(x \mid k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}) = \sum_{r=1}^{k} w_r Er(x \mid \nu_r, \mu_r), \qquad 0 < x < \infty, \tag{2.1}$$

donde $\sum_{r=1}^{k} w_r = 1$; $w_r, \mu_r > 0$, y $\nu_r \in \mathbb{N}$, para $r = 1, ..., k$, y donde,

$$Er(x \mid \nu_r, \mu_r) = \frac{(\nu_r \mu_r)^{\nu_r}}{\Gamma(\nu_r)} x^{\nu_r - 1} \exp(-\nu_r \mu_r x), \tag{2.2}$$

es la función de densidad de una distribución Erlang parametrizada para que su media sea $1/\mu_r$.

Una razón por la que se opta por esta clase de distribuciones es porque es densa en el conjunto de densidades definidas en $(0, \infty)$, (Teorema 6.2 Cap. III, Asmussen (1987)), es decir, para una distribución general, G, existe una sucesión de mixturas de Erlang que convergen débilmente a G. Otra ventaja, dentro del contexto de sistemas de colas, es que incluye las distribuciones utilizadas habitualmente, Erlang, hiperexponencial y exponencial, como casos particulares. Sin embargo, el motivo fundamental por el que se escogen mixturas de Erlang es que es una distribución continua de tipo fase (PH). La clase de distribuciones de tipo PH fue introducida por Neuts (1981) y se definen como la distribución del tiempo hasta la absorción en un proceso de Markov. En el Capítulo 3, se expone detalladamente su definición y algunas de sus características más importantes. Básicamente, las distribuciones de tipo PH son modelos que permiten descomponer cada observación en una sucesión de fases exponenciales. La Figura 2.1 ilustra una representación esquemática de las fases exponenciales involucradas en la distribución HEr. El hecho de que la mixtura de distribuciones Erlang sea de tipo PH permite aplicar los resultados de la Teoría de Colas clásica obtenidos por Neuts (1981) para sistemas de colas con distribución de servicio de tipo fase y tiempo entre llegadas exponencial que permitirán estimar mediante expresiones explícitas el número de clientes en el sistema en equilibrio, el tiempo de espera y la longitud del periodo de ocupación.

2.2.2. Distribución Coxiana (MGE).

Se introduce en esta Subsección a la familia de densidades obtenida a partir de mezclas de distribuciones Erlang generalizadas (MGE). Una distribución pertenece a la familia MGE si puede representarse como una

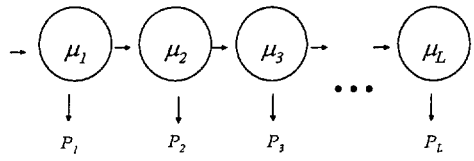


Figura 2.2: Representación de las fases exponenciales de la distribución MGE.

suma de exponenciales con tasas que pueden ser diferentes y donde el número de sumandos viene asignado por ciertas probabilidades previamente fijadas, véase la Figura 2.2. También, se trata de una distribución Coxiana en el sentido de Cox (1955), pero con tasas reales y probabilidades no negativas.

Sea X una variable aleatoria distribuida según una distribución MGE con parámetros $\mathbf{P} = (P_1, ..., P_L)$ y $\boldsymbol{\mu} = (\mu_1, ..., \mu_k)$. Entonces, la variable toma la forma siguiente,

$$X = \begin{cases} Y_1, & \text{con prob} = P_1 \\ Y_1 + Y_2, & \text{con prob} = P_2 \\ \vdots & \vdots \\ Y_1 + ... + Y_L, & \text{con prob} = P_L \end{cases} \tag{2.3}$$

donde $Y_r \sim \exp(\mu_r)$ y $\sum_{r=1}^{L} P_r = 1$. Por tanto, la correspondiente función de densidad de X se puede expresar como una mixtura,

$$f(x \mid L, \mathbf{P}, \boldsymbol{\mu}) = \sum_{r=1}^{L} P_r f_r(x \mid \mu_1, ..., \mu_r), \tag{2.4}$$

donde f_r es la función de densidad de una suma de r exponenciales, es decir, una distribución Erlang generalizada,

$$f_r(x \mid \mu_1, ..., \mu_r) = \sum_{t=1}^{r} \left(\prod_{s \neq t} \left(\frac{\mu_s - \mu_t}{\mu_s \mu_t} \right)^{-1} \right) \mu_t^{2-r} e^{-\mu_t x}, \tag{2.5}$$

véase, por ejemplo, Johnson y Kotz (1970). Obsérvese que en (2.5) se supone que todas las tasas, μ_r, $r = 1, ..., L$ son distintas. Se pueden obtener expresiones alternativas en caso de existir repetición de una o varias tasas.

Esta distribución es también de tipo PH e incluye la distribución exponencial ($L = 1$) y Erlang ($P_L = 1$, $\mu_1 = ... = \mu_L$), como casos particulares. Aunque no es evidente, esta familia de distribuciones contiene también a las mixturas de distribuciones Erlang, definidas anteriormente, y, en particular, a las mixturas de exponenciales. Esta propiedad se demostrará en el Capítulo siguiente y se comenta con más detalle en la siguiente Subsección. La distribución Coxiana es también densa en el conjunto de densidades continuas definidas en $(0, \infty)$, véase Bertsimas (1990). Y por tanto, incrementando el número de fases, L, es posible aproximar cualquier distribución continua definida en la recta real positiva utilizando una distribución MGE.

2.2.3. Comparación de las familias de distribuciones.

El conjunto de modelos de distribución que define la familia de mixturas de distribuciones Erlang está contenido en el conjunto de distribuciones definidas por el modelo MGE. Esta propiedad, mencionada anteriormente, se probará formalmente en el Capítulo 3 dentro del contexto de distribuciones de tipo PH,

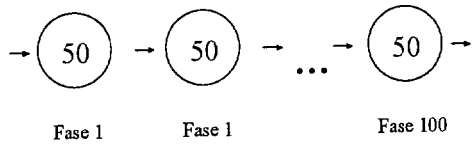


Figura 2.3: Representación de las fases exponenciales del ejemplo de distribución $Er\,(100, 0.5)$.

mostrándose cómo cada distribución HEr admite una representación propia del tipo MGE indicado en la Figura 2.2, que es propia de la distribución MGE.

Con este resultado podría pensarse que no es razonable desarrollar inferencia por separado para cada uno de los dos modelos de mixtura puesto que uno de ellos está contenido en el otro. Sin embargo, la parametrización es distinta en cada modelo y consecuentemente, existen diferencias importantes de cara a la inferencia en cuanto a flexibilidad y simplicidad. Para ilustrar este fenómeno, se puede considerar por ejemplo una muestra de datos procedentes de una distribución Erlang, $Er\,(100, 0.5)$, véase la Figura 2.3, que corresponde a la suma de 100 exponenciales de tasa 50. Con un método de estimación Bayesiano basado en una distribución HEr debería obtenerse como modelo más probable una mixtura con una única componente y valores medios a posteriori de ν y μ cercanos a 100 y 0.5, respectivamente. Sin embargo, con una distribución MGE, el modelo más probable debería ser una mixtura con un número muy elevado de componentes, próximo a 100, valores muy pequeños para los pesos, $\mathbf{P}$, excepto para el último que sería cercano a 1, y valores muy similares para las tasas, $\boldsymbol{\mu}$, alrededor de 50. Como se puede observar, en este caso, el número de parámetros necesarios a estimar con la distribución MGE es considerablemente superior al que necesita la mixtura HEr. Por tanto, el tiempo de computación requerido para la estimación será mucho mayor y el ajuste será más deficiente ya que la varianza a posteriori de la distribución de los parámetros será mayor. Además, en la práctica, no es frecuente, por motivos computacionales, asumir a priori que el número de componentes en la mixtura pueda tomar valores tan elevados.

Existen otras situaciones en las que el modelo HEr permite una aproximación más favorable utilizando menor cantidad de parámetros. Por ejemplo, para un conjunto de datos con múltiples modas, la distribución MGE necesitará un valor muy elevado del orden, L, para obtener una estimación satisfactoria a diferencia de la distribución HEr que, con un tamaño menor de la mixtura permitirá, generalmente, obtener un ajuste más adecuado en estos casos. Sin embargo, en muchas ocasiones, en las que aparecen datos asimétricos unimodales, será más adecuado ajustar una distribución MGE ya que es un modelo más general y por tanto, más flexible. Obsérvese que la distribución MGE consiste, básicamente, en sumas de exponenciales con tasas diferentes, mientras que la distribución HEr son sumas de exponenciales con la misma tasa. Además, para un valor fijo del tamaño de la mixtura, L ó k, el número de parámetros de una distribución MGE es menor que para el modelo HEr. Todos estos fenómenos se reflejarán en los ejemplos de la Sección 2.5. También, es reseñable que la distribución MGE da lugar a una mixtura con una ordenación natural de sus componentes. Esto es una ventaja con respecto a la inferencia ya que permite evitar problemas de intercambio de identificaciones causada por la simetría en la verosimilitud de los parámetros, véase Stephens (2000b). Y por tanto, no es necesario reparametrizar los valores de los parámetros del modelo como se hace en Mengersen y Robert (1996).

En términos generales, se puede afirmar que ambos modelos son apropiados para captar distribuciones con colas pesadas que son frecuentes en el servicio de algunos sistemas. Si bien este hecho facilita los cálculos de las distribuciones estacionarias de sistemas con este tipo de servicio, como se detalla en el Capítulo 3, es importante observar que las distribuciones HEr y MGE requieren, en general, más parámetros que la distribución Lognormal o Weibull y, consecuentemente, los procedimientos de estimación para las distribuciones HEr y MGE son más complejos.

Por último, el ajuste de un modelo de distribución MGE ofrece una ventaja adicional para la estimación en un sistema de colas ya que si se aproxima la distribución del tiempo de servicio y/o entre llegadas al sistema con un modelo MGE, además de permitir la obtención de expresiones para las distribuciones estacionarias, es más factible la obtención de las distribuciones transitorias del sistema, y por tanto, la estimación de las características de un sistema GI/G/1 en cualquier instante, τ, como se verá en el Capítulo 5.

2.3. Métodos MCMC asumiendo un número fijo de componentes.

Se pretende, ahora, hacer inferencia sobre los parámetros que rigen las distribuciones escogidas para los procesos de servicio y de llegadas. En esta Sección, se plantean dos procedimientos para ajustar los dos modelos de mixtura propuestos, HEr y MGE, a un conjunto de n observaciones, $\mathbf{x} = \{x_1, ..., x_n\}$, que representan los n_s tiempos de servicio, $\mathbf{s} = \{s_1, ..., s_{n_s}\}$, o los n_a tiempos entre las llegadas, $\mathbf{t} = \{t_1, ..., t_{n_a}\}$, a un determinado sistema de colas. Como se ha observado anteriormente, estos procedimientos se pueden utilizar también para representar cualquier variable con soporte en la semirecta real positiva y, especialmente, si presenta una cola asimétrica muy pesada, como se encuentra a menudo en tiempos de supervivencia.

Es fácil comprobar que no existe una distribución a priori conjugada para los parámetros, $\boldsymbol{\theta}$, de ninguno de los dos modelos de mixtura empleados y por tanto, no es sencillo calcular la expresión de la distribución a posteriori. Sin embargo, dada una distribución a priori, se puede desarrollar inferencia Bayesiana utilizando los métodos de cadenas de Markov Monte Carlo (MCMC). Básicamente, los métodos MCMC consisten en una serie de algoritmos que permiten aproximar integrales que no se pueden calcular analíticamente. En la Subsección 2.3.1, se presenta una breve recopilación de los fundamentos de los métodos MCMC y algunos de los algoritmos más comunes. Para una revisión extensa de esta materia, véase, por ejemplo, Gilks et al. (1996). En la Subsección 2.3.2, se describe un método MCMC para ajustar los datos a una mixtura de distribuciones Erlang y en la Subsección 2.3.3, a una distribución Coxiana. En esta Sección, se considera para los dos modelos que el número de componentes en la mixtura, k ó L, es conocido. En la Sección siguiente se incorporarán estos procedimientos en algoritmos MCMC más generales que permitan estimar conjuntamente el número de componentes y los parámetros asociados a cada una de ellas.

2.3.1. Métodos de Cadenas de Markov Monte Carlo (MCMC).

Los métodos MCMC permiten aproximar integrales que involucran una determinada distribución de interés que es desconocida. En un contexto Bayesiano, la distribución de interés es frecuentemente la distribución conjunta a posteriori, $f(\boldsymbol{\theta} \mid \mathbf{x})$, de los parámetros, $\boldsymbol{\theta}$, y se pretende habitualmente aproximar integrales que constituyen la esperanza a posteriori de una determinada función, $g(\boldsymbol{\theta})$, de los parámetros,

$$E[g(\boldsymbol{\theta}) \mid \mathbf{x}] = \int_{\Theta} g(\boldsymbol{\theta}) f(\boldsymbol{\theta} \mid \mathbf{x}) \, d\boldsymbol{\theta}, \tag{2.6}$$

donde Θ es el soporte de $\boldsymbol{\theta}$. La estrategia en los métodos MCMC es construir una cadena de Markov irreducible y aperiódica, $\{\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ...\}$ cuya distribución estacionaria, $\pi(\boldsymbol{\theta})$, coincida con la distribución a posteriori, $f(\boldsymbol{\theta} \mid \mathbf{x})$. Entonces,

$$\frac{1}{J} \sum_{j=1}^{J} g\left(\boldsymbol{\theta}^{(j)}\right) \overset{c.s.}{\rightarrow} E_{\pi}[g(\boldsymbol{\theta})], \qquad \text{cuando } J \rightarrow \infty.$$

Y por tanto, se pueden aproximar integrales del tipo (2.6) con las medias muestrales obtenidas a partir de muestras de la cadena de Markov en equilibrio. Evidentemente, si se pudiesen obtener muestras directamente de $f(\boldsymbol{\theta} \mid \mathbf{x})$ se podrían utilizar las técnicas de estimación Monte Carlo habituales, pero los métodos MCMC se utilizan cuando no es fácil generar valores de la distribución a posteriori de los parámetros. Cuando

se construye una cadena de Markov mediante un método MCMC se comienza con unos valores iniciales arbitrarios y se espera hasta que la cadena alcanza la convergencia. A continuación, como muestra de la distribución a posteriori se toman los siguientes valores generados por la cadena.

El problema en los métodos MCMC es encontrar la probabilidades de transición de la cadena que permitan que la distribución estacionaria sea la distribución de interés. Existen muchos métodos MCMC que ofrecen procedimientos para construir la cadena de Markov que se requiera. A menudo, estos procedimientos imponen una condición adicional para las probabilidades de transición que consiste en que la cadena de Markov sea reversible en el tiempo, lo que significa, básicamente, que la cadena se comporte de la misma manera si acontece hacia adelante o hacia atrás en el tiempo. Con esta condición, las ecuaciones de balance se verifican también para la distribución estacionaria, $\pi(\theta)$,

$$\pi(\theta)\, p\left(\theta, \bar{\theta}\right) = \pi\left(\bar{\theta}\right) p\left(\bar{\theta}, \theta\right), \qquad \text{para todo } \theta \text{ y } \bar{\theta},$$

donde $p\left(\theta, \bar{\theta}\right)$ es la probabilidad de transición de θ a $\bar{\theta}$. En general, es más sencillo obtener las probabilidades de transición a partir de estas ecuaciones que a partir de las ecuaciones de equilibrio, $\pi p = \pi$, y por esta razón, se suele imponer la condición de reversibilidad.

Los dos métodos MCMC más comunes son el muestreo de Gibbs, véase Gelfand y Smith (1990), y el método Metropolis Hastings, véase Hastings (1970). El muestreo de Gibbs se puede utilizar cuando es posible generar valores de la distribución a posteriori $f(\theta_i \mid \theta_{i-}, \mathbf{x})$ de cada parámetro, θ_i, condicionada al resto de parámetros $\theta_{i-} = \{\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_k\}$. Para generar valores de una cadena de Markov construida según un muestreo de Gibbs se comienza con unos valores iniciales arbitrarios, θ^0, y se permite que la cadena se mueva de θ^j a θ^{j+1} actualizando cada parámetro, θ_i^j, mediante un valor θ_i^{j+1} generado de $f\left(\theta_i \mid \theta_1^{j+1}, ..., \theta_{i-1}^{j+1}, \theta_{i+1}^j, ..., \theta_k^j, \mathbf{x}\right)$. Se puede demostrar que una cadena construida de esta manera tiene distribución estacionaria, $f(\theta \mid \mathbf{x})$, véase Smith y Roberts (1993).

El algoritmo Metropolis Hastings es un método MCMC alternativo que se puede usar cuando no se conoce la distribución a posteriori condicionada al resto de parámetros. Para generar valores de esta distribución lo que se propone es un movimiento candidato de θ^j a $\bar{\theta}^{j+1}$, generado de una distribución propuesta, $q\left(\theta, \bar{\theta}\right)$. El movimiento se acepta con probabilidad,

$$\alpha = \frac{f\left(\bar{\theta} \mid \mathbf{x}\right) q\left(\bar{\theta}, \theta\right)}{f(\theta \mid \mathbf{x})\, q\left(\theta, \bar{\theta}\right)}.$$

Si se acepta el movimiento los parámetros se actualizan estableciendo $\theta^{j+1} = \bar{\theta}^{j+1}$ y si se rechaza el movimiento se fija $\theta^{j+1} = \theta^j$. Esta probabilidad α constituye la probabilidad óptima de aceptar que verifica las condiciones de las ecuaciones de balance, véase Peskun (1973).

Existen muchos otros métodos MCMC y, en algunos casos, como se mostrará más adelante, se pueden construir algoritmos que combinen actualizaciones de tipo Metropolis con un muestreo de Gibbs, lo que se conoce como método híbrido Monte Carlo. En esta Sección, se utilizan métodos MCMC con dimensión fija de los parámetros, pero en la siguiente Sección se usarán otros procedimientos MCMC que permiten variar la dimensión paramétrica.

2.3.2. Inferencia para la distribución HEr.

En esta Subsección, se describe un algoritmo para hacer inferencia sobre los parámetros, $\theta = (\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu})$, de una mixtura de distribuciones Erlang, véase (2.1), con número de componentes, k, conocido. Inicialmente, del mismo modo que se hace en Diebolt y Robert (1994), se considera un conjunto de datos no observables

$\mathbf{z} = \{z_1, ..., z_n\}$ que indican la componente de la mixtura a la que pertenece cada una de las observaciones $\mathbf{x} = \{x_1, ..., x_n\}$. Estos datos faltantes son realizaciones de variables latentes i.i.d., $Z_1, ..., Z_n$, asociadas a cada dato y, como la proporción de observaciones procedentes de cada componente r viene dada por w_r, resulta natural asumir a priori,

$$P(Z_i = r \mid \mathbf{w}) = w_r, \qquad r = 1, ..., k. \tag{2.7}$$

De este modo, una vez observado el valor de las variables latentes, Z_i, las variables observables, X_i, son variables aleatorias independientes distribuidas según la densidad de su respectiva componente,

$$X_i \mid Z_i = r \sim Er(\nu_r, \mu_r), \qquad i = 1, ..., n. \tag{2.8}$$

Este *aumento de los datos* resulta muy adecuado para la interpretación del modelo y simplificación de los cálculos. Además, esta estructura de datos incompletos es muy frecuente en inferencia para modelos de mixturas, también fuera del contexto Bayesiano, véase Dempster et al. (1997), donde se hace uso del algoritmo EM.

En este contexto se define una distribución a priori para $(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{z})$, suponiendo que la distribución a priori conjunta se puede factorizar de la forma siguiente,

$$f(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{z}) = f(\mathbf{z} \mid \mathbf{w}) f(\mathbf{w}) f(\boldsymbol{\mu}) f(\boldsymbol{\nu}).$$

A continuación, se definen distribuciones a priori apropiadas para cada uno de los parámetros $\mathbf{w}$, $\boldsymbol{\mu}$ y $\boldsymbol{\nu}$. Es muy conocido que para modelos de mixturas no se pueden definir distribuciones a priori impropias e independientes para cada término de la mixtura ya que darían lugar a distribuciones a posteriori impropias originadas por la probabilidad positiva, que siempre existe, de que una componente cualquiera no genere ninguna observación, véase Diebolt y Robert (1994). Por tanto se definen distribuciones a priori propias pero muy poco informativas, al igual que, por ejemplo, en Wiper et al. (2001). Se considera,

$$\mathbf{w} \mid k \sim D(\phi_1, ..., \phi_k), \tag{2.9}$$

una distribución Dirichlet con parámetros $\phi_r > 0$. Como se hace habitualmente, se asigna $\phi_r = 1$, para todo $r = 1, ..., k$, con el objetivo de obtener una distribución uniforme para los pesos. En segundo lugar, se supone que la distribución a priori conjunta para $\boldsymbol{\mu}$, que son los valores inversos de las medias de las componentes de la mixtura, es proporcional a un producto de distribuciones gamma,

$$\mu_r \mid k \sim G(\alpha, \beta), \qquad \text{para } r = 1, ..., k, \tag{2.10}$$

con $\alpha, \beta > 0$ y restringido al subespacio $\mu_1 > ... > \mu_k$ por razones de identificabilidad. Hay que puntualizar que si el interés se centra únicamente en modelizar las variables observables, X, esta restricción no es necesaria. Como la distribución tiene que ser propia α y β tienen que ser distintos de 0. Además, si $\alpha \leq 1$ y $k > 1$ entonces los momentos de la distribución a priori de $1/\mu_r$, que se distribuye como una gamma invertida, no existen, lo cual implica que tampoco existirán los momentos de su distribución a posteriori. Consecuentemente, la media de la distribución predictiva de la variable de interés, X, asociada a la distribución del servicio o del tiempo entre llegadas, será infinita. Más aún, tampoco existirán los momentos de la intensidad de tráfico, ρ, véase (1.5), asociado al modelo de colas correspondiente. Se fija, por tanto, $\alpha = 1.1$ y $\beta = 1$ con objeto de obtener una distribución propia pero que no aporte demasiada información.

Finalmente, se considera una distribución geométrica (Geo) a priori con media $1/\vartheta$ para los parámetros enteros, ν_r. Y se fija, por ejemplo, $\vartheta = 0.01$, para obtener, como en los casos anteriores, una distribución a priori muy poco informativa. En la Sección 2.5, se incluye un análisis de sensibilidad con respecto a estas elecciones de los parámetros de las distribuciones a priori.

Obviamente, las distribuciones a priori que se acaban de exponer no son las únicas elecciones posibles y, de hecho, se han escogido principalmente por conveniencia. La formulación a priori para las variables indicadoras $\mathbf{Z}$ y los pesos $\mathbf{w}$ es la que se utiliza habitualmente en los modelos de mixturas, véase, por ejemplo, Diebolt y

Robert (1994). La distribución a priori para μ se ha seleccionado porque es semiconjugada y la distribución a priori geométrica para ν_r porque es un caso particular de la distribución binomial negativa (BN) y la forma de la distribución a posteriori de ν_r es parecida a esta familia como se verá más adelante.

Conocido el valor de k y dadas las distribuciones a priori que se acaban de exponer, es fácil calcular las distribuciones a posteriori condicionadas. Utilizando (2.7) y (2.8) se obtiene que,

$$P(Z_i = r \mid \mathbf{x}, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}) \propto w_r \frac{(\nu_r \mu_r)^{\nu_r}}{\Gamma(\nu_r)} x_i^{\nu_r - 1} \exp(-\nu_r \mu_r x_i), \qquad \text{para } r = 1, ..., k, \tag{2.11}$$

y además, a partir de (2.7), (2.9), (2.10) y los datos se tiene que,

$$\mathbf{w} \mid \mathbf{x}, \mathbf{z} \sim D(\phi_1 + n_1, ..., \phi_k + n_k), \tag{2.12}$$

$$\mu_r \mid \mathbf{x}, \mathbf{z} \sim G(\alpha + n_r \nu_r, \beta + S_r \nu_r), \tag{2.13}$$

donde $n_r = \#\{Z_i = r\}$, el número de observaciones que pertenecen a la componente r-ésima y $S_r = \sum_{i:Z_i=r} x_i$, la suma de los valores de estas observaciones, para $r = 1, ..., k$. Finalmente, como,

$$f(\nu_r \mid \mathbf{x}, \mathbf{z}, \mathbf{w}, \boldsymbol{\mu}) \propto \left[\prod_{i:Z_i=r} \frac{(\nu_r \mu_r)^{\nu_r}}{\Gamma(\nu_r)} x_i^{\nu_r - 1} \exp(-\nu_r \mu_r x_i) \right] \times \left[\vartheta (1-\vartheta)^{\nu_r - 1} \right],$$

entonces,

$$f(\nu_r \mid \mathbf{x}, \mathbf{z}, \mathbf{w}, \boldsymbol{\mu}) \propto \frac{\nu_r^{n_r \nu_r}}{\Gamma(\nu_r)^{n_r}} \exp\left\{ -\nu_r \left(-\log(1-\vartheta) + S_r \mu_r - n_r \log \mu_r - \log P_r \right) \right\}, \tag{2.14}$$

donde $P_r = \prod_{i:Z_i=r} x_i$ es el producto de las observaciones que pertenecen a la componente r.

Ahora, se puede definir un algoritmo MCMC para obtener muestras de la distribución a posteriori de los parámetros $(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{z})$ en el que el número de términos en la mixtura, k, está fijo. Para ello, se generan valores de una cadena de Markov irreducible y aperiódica cuya distribución estacionaria es la distribución a posteriori de los parámetros. El algoritmo consiste en un muestreo de Gibbs, véase Gelfand y Smith (1990), modificado con pasos Metropolis Hastings, véase Hastings (1970). El esquema es el siguiente y los pasos más importantes se detallan a continuación.

ALGORITMO HEr.

1. Fijar valores iniciales para $\mathbf{w}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\nu}^{(0)}$.
2. Actualizar los datos faltantes generando de $\mathbf{z}^{(j+1)} \sim \mathbf{z}|\mathbf{x}, \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\nu}^{(j)}$.
3. Actualizar los pesos generando de $\mathbf{w}^{(j+1)} \sim \mathbf{w}|\mathbf{x}, \mathbf{z}^{(j+1)}$.
4. Para $r = 1, ..., k$,
   a. Actualizar μ_r generando de $\mu_r^{(j+1)} \sim \mu_r|\mathbf{x}, \mathbf{z}^{(j+1)}$.
   b. Actualizar ν_r utilizando un paso Metropolis Hastings.
5. Ordenar $\boldsymbol{\mu}^{(j+1)}$ y colocar $\mathbf{w}^{(j+1)}$ y $\boldsymbol{\nu}^{(j+1)}$ según este orden.
6. $j = j + 1$. Ir a 2.

En el paso 1, se escogen valores iniciales para los parámetros de la mixtura. En los pasos 2, 3 y 4a, se generan valores de la distribución condicionada a posteriori de $\mathbf{z}, \mathbf{w}$ y $\boldsymbol{\mu}$, que tienen todos distribuciones comunes y fáciles de muestrear dadas por (2.11), (2.12) y (2.13), respectivamente. El paso complicado es el

4b, en el que hay que generar valores de la distribución de $f(\nu_r|\,\mathbf{x},\mathbf{z},\mathbf{w},\boldsymbol{\mu})$, para $r=1,...,k$. La constante de proporcionalidad es igual a la suma de la expresión dada en (2.14) para todos los valores de $\nu_j=1,...,\infty$. Como se trata de una variable discreta se podrían evaluar los sumandos hasta aproximar suficientemente la suma total. Sin embargo, este procedimiento es muy lento computacionalmente. Por eso, se introduce un método Metropolis Hasting para generar valores de la distribución a posteriori de $\boldsymbol{\nu}$. Se generan valores candidatos $\tilde{\nu}_r$ utilizando como distribución propuesta una binomial negativa,

$$f_{BN}(\nu)=\binom{m+\nu-2}{\nu-1}p^m(1-p)^{\nu-1},\qquad \nu=1,2,... \tag{2.15}$$

porque para valores grandes de ν_r, la forma de la distribución a posteriori condicionada obtenida en (2.14) es parecida a una gamma y la versión discreta de ésta es la binomial negativa. El candidato, $\tilde{\nu}_r$, se acepta con probabilidad $\alpha=\min\{1,A_\nu\}$, donde ,

$$A_\nu=\frac{f(\tilde{\nu}_r\mid \mathbf{x},\mathbf{z}^{(j+1)},\mathbf{w}^{(j+1)},\boldsymbol{\mu}^{(j+1)})p(\tilde{\nu}_r,\nu_r^{(j)})}{f(\nu_r^{(j)}\mid \mathbf{x},\mathbf{z}^{(j+1)},\mathbf{w}^{(j+1)},\boldsymbol{\mu}^{(j+1)})p(\nu_r^{(j)},\tilde{\nu}_r)}, \tag{2.16}$$

y donde $p(\nu_r^{(j)},\tilde{\nu}_r)$ es la probabilidad de generar $\tilde{\nu}_r$ dado el valor de la iteración anterior $\nu_r^{(j)}$.

Los valores de los parámetros, (m,p), de la distribución propuesta en 2.15 se suelen seleccionar de modo que la media sea el valor $\nu_r^{(j)}$ de la iteración anterior. Sin embargo, en este caso aparece un problema si $\nu_r^{(j)}=1$, porque entonces se genera el candidato $\tilde{\nu}_r=1$ con probabilidad uno. Ante esta situación, se tiene una cadena reducible ya que se acepta siempre el valor generado y la cadena no se mueve para los valores de ν_j. Aunque esta dificultad se puede solucionar modificando ligeramente los parámetros, después de examinar varias situaciones, se opta por una alternativa que consiste en elegir los parámetros m y p de la distribución propuesta de manera que su moda sea el valor $\nu_r^{(j)}$ de la iteración anterior. Se puede comprobar que esta condición se verifica si,

$$\frac{m-1}{\nu_r^{(j)}+m-1}<p<\frac{m-1}{\nu_r^{(j)}+m-2}.$$

Por ejemplo, en la práctica se ha fijado $m=\nu_r^{(j)}+1$ y $p=(m-1)/(\nu_r^{(j)}+m-1.5)$ para obtener una varianza que de lugar a una tasa razonable de valores aceptados.

Consecuentemente, la expresión para A_ν, dada en (2.16), es igual a,

$$\begin{aligned}A_\nu &= \left(\frac{\Gamma(\nu_r^{(j)})\tilde{\nu}_r^{\tilde{\nu}_r}}{\Gamma(\tilde{\nu}_r)\nu_r^{(j)\nu_r^{(j)}}}\right)^{n_r}\exp\left\{\left(\nu_r^{(j)}-\tilde{\nu}_r\right)\times\left(-\log(1-\vartheta)+S_r\mu_r^{(j+1)}-n_r\log\mu_r^{(j+1)}-\log P_r\right)\right\}\times\\ &\quad\times\frac{\tilde{\nu}_r}{\nu_r^{(j)}}\left(\frac{\nu_r^{(j)}}{2\nu_r^{(j)}-0.5}\right)^{\nu_r^{(j)}+1}\left(\frac{2\tilde{\nu}_r-0.5}{\tilde{\nu}_r}\right)^{\tilde{\nu}+1}\left(\frac{\nu_r^{(j)}-0.5}{2\nu_r^{(j)}-0.5}\right)\left(\frac{2\tilde{\nu}_r-0.5}{\tilde{\nu}_r-0.5}\right)^{\nu_r^{(j)}-1}.\end{aligned}$$

Se sabe que para una cadena como la que se ha construido a partir del algoritmo `HEr`, las medias convergen a su esperanza bajo la distribución estacionaria, π, de la cadena (si existen y son finitas), véase la Susbsección 2.3.1,

$$\bar{\theta}_r=\frac{1}{J-B}\sum_{j=B+1}^{J+B}\theta_r^{(j)}\overset{c.s.}{\to}E_\pi[\theta_r], \tag{2.17}$$

donde $\theta_r=w_r,\mu_r,\nu_r$ para $r=1,2,...k$, J el tamaño de la muestra Monte Carlo y B el número necesario de iteraciones para tener convergencia (*burn-in*). Por tanto, las medias ergódicas, $\bar{\theta}_r$, constituyen estimadores consistentes de las esperanzas de la distribución conjunta a posteriori de $\boldsymbol{\theta}$ que coincide con la distribución

estacionaria de la cadena, π. Además, dada una realización de la cadena de tamaño J, la distribución predictiva de la variable de interés, X, se puede estimar con la expresión siguiente. Esta estimación está basada en el teorema de Rao-Blackwell, véase Casella y Robert (1996), y es muy habitual en la aproximación de integrales a partir de muestras MCMC.

$$f(x \mid \mathbf{x}) = \frac{1}{J} \sum_{j=B+1}^{J+B} \sum_{r=1}^{k} w_r^{(j)} Er(x \mid \nu_r^{(j)}, \mu_r^{(j)}). \tag{2.18}$$

2.3.3. Inferencia para la distribución MGE.

En esta Subsección, el punto de partida es un conjunto de observaciones independientes $\mathbf{x} = \{x_1, ..., x_n\}$ que se suponen generadas por una distribución MGE, véase (2.3), de orden, L, conocido. El objetivo es construir un algoritmo MCMC para obtener una muestra de la distribución a posteriori de los parámetros del modelo $\boldsymbol{\theta} = (\mathbf{P}, \boldsymbol{\mu})$ y poder desarrollar inferencia Bayesiana sobre los parámetros y la variable de interés, X.

Al igual que en la Sección anterior, se introduce una estructura de datos faltantes, algo más sofisticada, que permite simplificar considerablemente la función de verosimilitud. Se incluye, en primer lugar, una variable, Z_i, que informa sobre la componente de la mixtura a la que pertenece cada variable observable, X_i, que también, puede interpretarse como la fase en la que cada observación abandona la serie de estados que constituye el modelo MGE, véase la Figura 2.2. En segundo lugar, se define un conjunto de variables, Y_{ir}, que proporciona el tiempo que permanece la observación i-ésima en el estado r, para $r = 1, ..., Z_i$, de modo que $\sum_{r=1}^{Z_i} Y_{ir} = X_i$. Entonces, se obtiene,

$$f(z_i, y_{i1}, ..., y_{iz_i} \mid \boldsymbol{\theta}) = P_{z_i} \prod_{r=1}^{z_i} \mu_r \exp(-\mu_r y_{ir}),$$

donde $\boldsymbol{\theta} = (\mathbf{P}, \boldsymbol{\mu})$. Se denota por $\mathbf{z} = (z_i)$ y por $\mathbf{y} = (y_{ir})$ al vector y a la matriz de datos no observables, respectivamente, para $i = 1, ..., n$ y para $r = 1, ..., z_i$. Obsérvese que el conjunto de datos faltantes, $(\mathbf{z}, \mathbf{y})$, proporciona por sí solo información completa sobre los datos de modo que conocido su valor se puede reconstruir el valor de las observaciones $\mathbf{x}$.

Para desarrollar inferencia Bayesiana, es necesario definir distribuciones a priori para los parámetros, $\boldsymbol{\theta}$, del modelo. Se supone a priori que los pesos, $\mathbf{P}$, y las tasas, $\boldsymbol{\mu}$, son independientes. Conocido el valor de L, se asumen distribuciones a priori semiconjugadas para $(\mathbf{P}, \boldsymbol{\mu})$. Se define una distribución Dirichlet para los pesos, $\mathbf{P} \sim D(1, ...1)$. Y para el caso de las tasas, $\boldsymbol{\mu}$, se añade un estrato más en la jerarquía permitiendo que su distribución a priori dependa de un hiperparámetro η. En concreto, se consideran distribuciones exponenciales con tasa común,

$$\mu_r \mid \eta \sim Exp(\eta), \qquad \text{para } r = 1, ..., L,$$

y una distribución gamma a priori,

$$\eta \sim G(g, h).$$

Se fijan $g = 0$ y $h = 0$ de modo que se obtiene una distribución impropia a priori para η. Esta estructura jerárquica permite que las tasas, $\boldsymbol{\mu}$, se estimen enteramente a partir de los datos. Las distribuciones a priori que se acaban de definir han sido seleccionadas, principalmente, por conveniencia en los cálculos ya que se trata de distribuciones semiconjugadas. Sin embargo, es evidente que se pueden considerar otras posibilidades.

Se exponen seguidamente las distribuciones a posteriori de los parámetros del modelo condicionadas al resto de parámetros. Estas distribuciones van a permitir la construcción de una cadena MCMC consistente en un muestreo de Gibbs. Obsérvese que, dentro de un entorno Bayesiano, las variables faltantes representan

un conjunto adicional de parámetros. Dadas las variables latentes, Z_i y Y_{ir}, es fácil probar que la distribución a posteriori condicionada de $\mathbf{P}$ es también una distribución Dirichlet y las tasas, μ_r, siguen distribuciones Gamma,

$$\mathbf{P} \mid \mathbf{x}, \mathbf{z}, \mathbf{y} \sim D(1+n_1, ..., 1+n_L), \tag{2.19}$$

donde $n_r = \#\{Z_i = r\}$ y

$$\mu_r \mid \mathbf{x}, \mathbf{z}, \mathbf{y}, \eta \sim G(1+n'_r, \eta + S'_r), \tag{2.20}$$

donde $n'_r = \#\{Z_i \geq r\}$ y $S'_r = \sum_{i:Z_i \geq r} y_{ir}$. Obsérvese que la información suficiente que aportan los datos faltantes, $(\mathbf{z}, \mathbf{y})$, consta de tres elementos: el número observaciones que abandonan en la fase r-ésima, n_r, el número de observaciones que atraviesan esa fase, n'_r, y el tiempo total que se permanece en ella, S'_r. Por otro lado, se puede comprobar fácilmente que la distribución a posteriori del hiperparámetro es,

$$\eta \mid \mathbf{x}, \boldsymbol{\mu} \sim G\left(g+L, h+\Sigma\mu_r\right). \tag{2.21}$$

Las variables latentes se pueden generar obteniendo muestras de los dos factores de

$$f\left(z_i, y_{i1}, ..., y_{iz_i} \mid x_i, \theta\right) = f\left(y_{i1}, ..., y_{iz_i} \mid x_i, z_i, \theta\right) f\left(z_i \mid x_i, \theta\right),$$

para $i = 1, ..., n$. La distribución a posteriori condicional $f\left(z_i \mid x_i, \theta\right)$ se puede obtener a partir de (2.5) y viene dada por,

$$P\left(z_i = r \mid x_i, \theta\right) \propto P_r \sum_{t=1}^{r} \left(\prod_{s \neq t} \left(\frac{\mu_s - \mu_r}{\mu_s \mu_r} \right)^{-1} \right) \mu_t^{2-r} e^{-\mu_t x_i}. \tag{2.22}$$

Es un poco más complicado generar valores de $f\left(y_{i1}, ..., y_{iz_i} \mid x_i, z_i, \theta\right)$ que es la distribución conjunta de los tiempos que permanece la observación x_i en cada uno de los estados condicionado a que abandona en el estado z_i. Se puede comprobar que esta distribución es un producto de z_i exponenciales restringidas al subespacio $\sum_{r=1}^{z_i} y_{ir} = x_i$. Asumiendo, sin pérdida de generalidad, que $\min\{\mu_r\} = \mu_s$ y suponiendo que $\mu_r \neq \mu_s$ para todo $r \neq s$, se obtiene,

$$f\left(y_{i1}, ..., y_{i,s-1}, y_{i,s+1}, ..., y_{i,z_i} \mid x_i, z_i, \theta\right) \propto \prod_{\substack{r=1 \\ r \neq s}}^{z_i} \left(\mu_r - \mu_s\right) \exp\left\{-\left(\mu_r - \mu_s\right) y_{ir}\right\}, \tag{2.23}$$

definido sobre la región $\mathbf{R} = \{y_{i1} + ... + y_{i,s-1} + y_{i,s+1} + ... + y_{i,z_i} \leq x_i\}$. Aunque se pueden obtener expresiones alternativas a (2.23) para el caso en el que $\boldsymbol{\mu}$ tiene varios mínimos, este cálculo no es necesario ya que dentro de un muestreo de Gibbs los valores de $\boldsymbol{\mu}$ son valores generados de distribuciones gamma que, teóricamente, tienen probabilidad cero de dar lugar a dos tasas exactamente iguales.

Con las distribuciones que se acaban de describir, se puede construir una cadena de Markov cuya distribución estacionaria sea la distribución a posteriori de los parámetros. El algoritmo MCMC consiste, fundamentalmente, en un muestreo de Gibbs, véase Gelfand y Smith (1990), en el que se incluye un aumento de datos mediante las variables latentes. La estructura del algoritmo es la siguiente:

ALGORITMO MGE.

1. Fijar valores iniciales $\theta^{(0)} = \left(\mathbf{P}^{(0)}, \boldsymbol{\mu}^{(0)}\right)$.
2. Actualizar las variables latentes.
   - 2.1. Generar valores de $\mathbf{z}^{(j+1)} \sim \mathbf{z} \mid \mathbf{x}, \mathbf{P}^{(j)}, \boldsymbol{\mu}^{(j)}$.
   - 2.2. Generar valores de $\mathbf{y}^{(j+1)} \sim \mathbf{y} \mid \mathbf{x}, \mathbf{z}^{(j+1)}, \mathbf{P}^{(j)}, \boldsymbol{\mu}^{(j)}$.
3. Actualizar el valor del hiperparámetro $\eta^{(j+1)} \sim \eta \mid \boldsymbol{\mu}^{(j)}$.

4. Actualizar los parámetros específicos.

   4.1. Generar valores de los pesos $\mathbf{P}^{(j+1)} \sim \mathbf{P}|\ \mathbf{x},\mathbf{z}^{(j+1)},\mathbf{y}^{(j+1)}$.

   4.2. Generar valores de las tasas $\boldsymbol{\mu}^{(j+1)} \sim \boldsymbol{\mu}|\ \mathbf{x},\mathbf{z}^{(j+1)},\mathbf{y}^{(j+1)}$.

5. $j = j+1$. Ir a 2.

Este algoritmo es fácil de implementar ya que las distribuciones condicionadas son explícitas. En el paso 2 se generan los datos faltantes $(\mathbf{z},\mathbf{y})$ a partir de las distribuciones dadas en (2.22) y (2.23). Teniendo en cuenta que la función de densidad dada en (2.23) es una exponencial multivariante truncada en la región $\mathbf{R}$, se pueden obtener muestras utilizando, por ejemplo, el método de rechazo, véase Ripley (1987), generando candidatos de una exponencial multivariante o de una uniforme definida en $\mathbf{R}$.

En el paso 3, se genera un valor de la distribución condicionada a posteriori de η dada en (2.21) y finalmente, en el paso 4, se generan valores de condicionada a posteriori de $\mathbf{P}$ y $\boldsymbol{\mu}$, dadas en (2.19) y (2.20), respectivamente.

Al igual que en la Sección anterior, se puede hacer inferencia sobre los parámetros de la distribución MGE utilizando las muestras obtenidas a partir de la cadena de Markov que constituye el algoritmo descrito. Por ejemplo, se pueden estimar las medias a posteriori de los parámetros con la correspondientes medias muestrales de la muestra Monte Carlo, como se indica en (2.17). También es posible aproximar la distribución predictiva de la variable de interés, X, con una expresión análoga a la que se muestra en (2.18), utilizando,

$$f(x \mid \mathbf{x}) = \frac{1}{J} \sum_{j=B+1}^{J+B} \sum_{r=1}^{L} P_r^{(j)} f_r \left(x \mid \mu_1^{(j)}, ..., \mu_r^{(j)} \right),$$

donde f_r es la densidad de una Erlang generalizada que corresponde a la suma de r exponenciales de tasas $\left(\mu_1^{(j)}, ..., \mu_r^{(j)}\right)$ dada en (2.5). Y donde L es el número fijo de componentes en la distribución MGE, J el tamaño de la muestra Monte Carlo y B el número necesario de iteraciones para tener convergencia (*burn-in*).

2.4. Métodos de dimensión paramétrica variable.

En esta Sección, se supone que el número de términos que intervienen en los modelos de mixtura es desconocido. Se pretende, como antes, ajustar un conjunto de observaciones $\mathbf{x} = \{x_1, ..., x_n\}$ a uno de los dos modelos de mixtura sugeridos, HEr y MGE, suponiendo, además, que no hay información sobre el número de grupos homogéneos, k ó L. Se proponen extensiones de los algoritmos MCMC expuestos en la Sección anterior construyendo cadenas de Markov que puedan moverse sobre un espacio de parámetros de dimensión variable. Para ello, se hace uso de dos metodologías diferentes que permiten hacer inferencia sobre el tamaño de la mixtura y, simultáneamente, sobre el resto de parámetros. De una parte, los métodos de salto reversible (RJMCMC) introducidos por Richardson y Green (1997) y por otro lado, los métodos basados en procesos de nacimiento y muerte (BDMCMC) introducidos por Stephens (2000a). La diferencia fundamental entre ambos tipos de algoritmos es que los métodos BDMCMC se basan en la construcción de procesos de Markov que acontecen en tiempo continuo a diferencia de la técnica del salto reversible que consiste, brevemente, en una extensión de un método Metropolis Hastings adaptado a un espacio de estados de dimensión variable.

En primer lugar, es necesario definir una distribución a priori discreta para el tamaño de la mixtura, k ó L. Una posibilidad es una uniforme discreta definida en el intervalo cerrado comprendido entre 1 hasta un valor máximo determinado, por ejemplo, 10. Alternativamente, se puede considerar una distribución Poisson con media pequeña y truncada en el intervalo $[1, 10]$. Esta elección tiene la ventaja de penalizar la sobreparametrización atribuyendo a priori una probabilidad pequeña para valores altos del número de

términos en la mixtura. Se pueden considerar muchas otras elecciones a priori para la dimensión, k ó L. En la Sección 2.5, se analiza la sensibilidad a estas elecciones sobre la distribución a posteriori de k ó L.

Dada una distribución a priori sobre el número de componentes en la mixtura, k ó L, todos los métodos de cadenas de Markov Monte Carlo que exploran modelos de mixturas de dimensión variable se basan en la construcción de cadenas que permiten movimientos a lo largo de la distribución a posteriori de k ó L. Para ello, se definen movimientos específicos entre modelos de distribución con diferente dimensión paramétrica que acontecen en tiempo discreto o continuo, según el tipo de algoritmo. Los movimientos de la cadena a lo largo de subespacios con dimensión paramétrica constante se llevan a cabo incorporando los algoritmos `HEr` y `MGE` descritos en la Sección anterior.

En la Subsección 2.4.1, se describen dos métodos de tipo RJMCMC para hacer inferencia sobre los dos modelos de mixtura, HEr y MGE. Y en la Subsección 2.4.2 se proponen, alternativamente, dos algoritmos del tipo BDMCMC con el propósito de desarrollar comparaciones entre ambas metodologías. En la Tabla 2.1, se muestra un esquema que incluye los dos algoritmos desarrollados en la Sección anterior y los que se describen en esta Sección.

	Modelo de mixtura	
	Distribución HEr	Distribución MGE
Dimensión paramétrica fija:	Algoritmo HEr	Algoritmo MGE
Dimensión variable con RJMCMC:	Algoritmo RJHEr	Algoritmo RJMGE
Dimensión variable con BDMCMC:	Algoritmo BDHEr	Algoritmo BDMGE

Tabla 2.1: Esquema de los algoritmos que se desarrollan en este Capítulo.

2.4.1. Métodos de salto reversible (RJMCMC).

Se proponen a continuación dos algoritmos del tipo salto reversible para las dos distribuciones HEr y MGE con la misma finalidad del apartado anterior, es decir, construir una cadena de Markov irreducible y aperiódica cuya distribución estacionaria coincida con la distribución conjunta a posteriori de los parámetros de los modelos de mixtura, pero incluyendo en este apartado el número de términos de la misma. Como en los métodos Metropolis Hastings, en los métodos RJMCMC, se proponen movimientos candidatos para el número de componentes en la mixtura que se aceptan o rechazan con una probabilidad que permite que la distribución de los valores aceptados satisfaga las propiedades de reversibilidad. De la misma manera que Richardson y Green (1997), se plantean dos tipos de movimientos que proponen cambiar el número de términos en una unidad. En concreto, estos movimientos son la separación o combinación de componentes de la mixtura, y el nacimiento o muerte de componentes vacías.

2.4.1.1. Método RJMCMC para ajustar una distribución HEr.

El procedimiento del algoritmo es el siguiente. Para un valor fijo de k, se utiliza el algoritmo `HEr` para generar valores del resto de parámetros. Entonces, se hace uso del método del salto reversible para generar valores de la distribución condicionada a posteriori de k y modificar los demás parámetros coherentemente. Se expone a continuación el esquema del algoritmo que se ilustra de manera gráfica en la Figura 2.4. Los pasos 3 y 4, que son los que modifican la dimensión de los parámetros, se especifican seguidamente.

`ALGORITMO RJHEr.`

1. `Fijar valores iniciales para` $k^{(0)}, \mathbf{w}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\nu}^{(0)}$.

2. `ALGORITMO HEr: Generar valores de` $\left(\mathbf{z}^{(j+1)}, \mathbf{w}^{(j+1)}, \boldsymbol{\mu}^{(j+1)}, \boldsymbol{\nu}^{(j+1)}\right)$ `para` $k^{(j)}$ `fijo.`

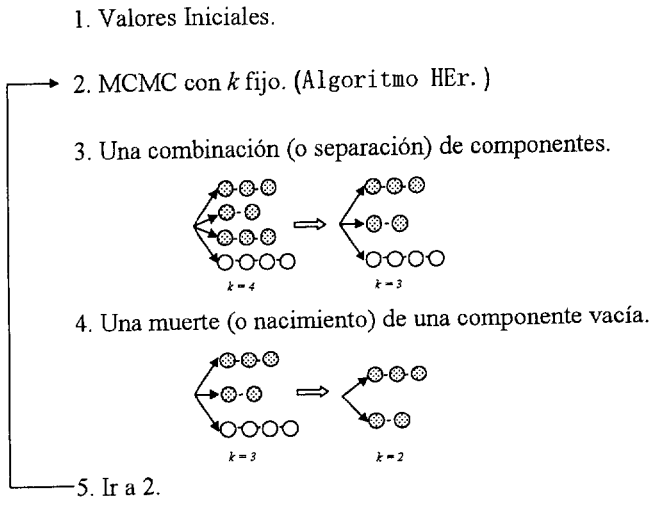


Figura 2.4: Algoritmo RJHEr asociado al modelo de mixtura HEr y basado en las técnicas de salto reversible. Las fases en blanco representan componentes vacías y en gris, fases con datos asociados.

3. Separar una componente de la mixtura en dos, o combinar dos en una.

4. Nacimiento o muerte de una componente vacía.

5. $j = j + 1$. Ir a 2.

En los pasos 3 y 4, se introduce un salto reversible para permitir que la cadena se mueva dentro de los valores de la distribución a posteriori del tamaño de la mixtura, k. Se proponen valores candidatos para k y se modifica el resto de los parámetros para adecuarlos a la nueva dimensión. En primer lugar, en el paso 3, se genera un candidato para el número de componentes de la mixtura, $\tilde{k}$, decidiéndose aleatoriamente entre separar una componente en dos o combinar dos componentes en una.

En caso de un movimiento de combinación, se eligen al azar dos componentes adyacentes (r_1, r_2) para fusionarse en una de modo que se reduce el valor de k en una unidad. Además, se modifica el resto de los parámetros del siguiente modo,

1. $\tilde{w} = w_{r_1} + w_{r_2}$.

2. $\frac{1}{\tilde{\mu}} = \frac{w_{r_1}}{\tilde{w}\mu_{r_1}} + \frac{w_{r_2}}{\tilde{w}\mu_{r_2}}$.

3. $\tilde{\nu} = \nu_{r_1}$.

Con estas transformaciones, se preservan los momentos de orden 0 y 1 de la distribución que se quiere estimar. Obsérvese que las transformaciones que se han usado para definir $\tilde{w}$ y $\tilde{\mu}$ son análogas a las empleadas en Richardson y Green (1997). Y se ha elegido la fórmula para $\tilde{\nu}$ por simplicidad, puesto que debido al soporte discreto de ν no se puede preservar el momento de orden 2 manteniendo la reversibilidad, como se comenta más adelante. Obviamente, se pueden utilizar muchas otras expresiones. Además, todas las observaciones procedentes de cualquiera de las dos componentes a combinar, es decir, con $z_i = r_1, r_2$, se asignan a la nueva componente, $\tilde{z}_i = r$.

En caso de un movimiento de separación, se elige al azar una componente, r, para dividirse en dos. Para

obtener los nuevos parámetros, se generan dos valores, u_1 y u_2 de una distribución uniforme $U(0,1)$ y un tercer valor, u_3, de una binomial negativa con moda ν_r. Entonces, se modifican los demás parámetros de la manera siguiente,

1. $\tilde{w}_{r_1} = u_1 w_r, \qquad \tilde{w}_{r_2} = (1-u_1) w_r.$

2. $\frac{1}{\tilde{\mu}_{r_1}} = \frac{1}{\mu_{r-1}} + u_2 \left(\frac{1}{\mu_r} - \frac{1}{\mu_{r-1}} \right), \qquad \frac{1}{\tilde{\mu}_{r_2}} = \frac{1-u_1 u_2}{(1-u_1)\mu_r} - \frac{u_1(1-u_2)}{\mu_{r-1}}.$

3. $\tilde{\nu}_{r_1} = \nu_r, \qquad \tilde{\nu}_{r_2} = u_3.$

Se han usado estas transformaciones inversas para mantener la reversibilidad. Esto significa que, si por ejemplo, se ha propuesto un candidato a partir de la separación de dos componentes debe de existir una probabilidad positiva de obtener los valores originales de los parámetros si se aplica un movimiento de combinación sobre ellos. La elección para los candidatos a pesos, $\tilde{w}_{r_i}$, es análoga a la empleada en Richardson y Green (1997), y para las inversas de las medias, $\tilde{\mu}_{r_i}$, es similar a la utilizada en Wiper et al. (2001). La expresión para $\tilde{\nu}$ se ha elegido de manera que sea lo más simple posible manteniendo la reversibilidad. Además, las observaciones procedentes de la componente a combinar, es decir con $z_i = r$, se asignan a cada una de las dos nuevas componentes, $\tilde{z}_i = r_1$ ó $\tilde{z}_i = r_2$, con probabilidades que son proporcionales a la expresión dada en (2.11) y vienen dadas por

$$P(\tilde{z}_i = r_j) = \frac{\tilde{w}_{r_j} Er(x_i \mid \tilde{\nu}_{r_j}, \tilde{\mu}_{r_j})}{\tilde{w}_{r_1} Er(x_i \mid \tilde{\nu}_{r_1}, \tilde{\mu}_{r_1}) + \tilde{w}_{r_2} Er(x_i \mid \tilde{\nu}_{r_2}, \tilde{\mu}_{r_2})}, \qquad j = 1, 2. \tag{2.24}$$

Por último, se acepta el movimiento con probabilidad mín $\{1, A_k^{sc}\}$, donde,

$$A_k^{sc} = \frac{f(\bar{\theta} \mid \mathbf{x}) p(\bar{\theta}, \theta)}{f(\theta \mid \mathbf{x}) p(\theta, \bar{\theta})}, \tag{2.25}$$

y donde $f(\theta \mid \mathbf{x})$ es la distribución a posteriori de los parámetros y $p(\theta, \bar{\theta})$ es la probabilidad de moverse desde $\theta = (k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu})$ a $\bar{\theta} = (\tilde{k}, \tilde{\mathbf{w}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}})$. En concreto, si $\tilde{k} = k+1$, se tiene por un lado que,

$$\frac{f(\bar{\theta} \mid \mathbf{x})}{f(\theta \mid \mathbf{x})} = \frac{f\left(\tilde{k}\right) f\left(\tilde{\mathbf{w}} \mid \tilde{k}\right) f\left(\tilde{\mathbf{z}} \mid \tilde{k}, \tilde{\mathbf{w}}\right) f\left(\tilde{\boldsymbol{\mu}} \mid \tilde{k}\right) f\left(\tilde{\boldsymbol{\nu}} \mid \tilde{k}\right) f\left(\mathbf{x} \mid \tilde{\mathbf{z}}, \tilde{k}, \tilde{\mathbf{w}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\nu}}\right)}{f(k) f(\mathbf{w} \mid k) f(\mathbf{z} \mid k, \mathbf{w}) f(\boldsymbol{\mu} \mid k) f(\boldsymbol{\nu} \mid k) f(\mathbf{x} \mid \mathbf{z}, k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu})},$$

por tanto, se obtiene,

$$\begin{aligned} \frac{f(\bar{\theta} \mid \mathbf{x})}{f(\theta \mid \mathbf{x})} &= \frac{f(k+1)}{f(k)} \times \frac{\Gamma(k+1)}{\Gamma(k)} \times \frac{\tilde{w}_{r_1}^{\tilde{n}_{r_1}} \tilde{w}_{r_2}^{\tilde{n}_{r_2}}}{w_r^{n_r}} \times \frac{(k+1)\, \tilde{\mu}_{r_1}^{-0.1} e^{-1/\tilde{\mu}_{r_1}} \tilde{\mu}_{r_2}^{-0.1} e^{-1/\tilde{\mu}_{r_2}}}{\mu_r^{-0.1} e^{-1/\mu_r}} \times \\ &\times (1-\vartheta)^{\tilde{\nu}_{r_1} + \tilde{\nu}_{r_2} - \nu_r} \times \frac{\prod_{i:\tilde{z}_i = r_1, r_2} Er(x_i \mid \tilde{\nu}_{\tilde{z}_i}, \tilde{\mu}_{\tilde{z}_i})}{\prod_{i:z_i = r} Er(x_i \mid \nu_r, \mu_r)}, \end{aligned}$$

y, por otra parte,

$$\frac{p(\bar{\theta}, \theta)}{p(\theta, \bar{\theta})} = \frac{d_{k+1}}{b_k \prod_{i:z_i = r} P_{\tilde{z}_i}} \times \frac{1}{f_{BN}(u_3)} \times \frac{w_r^2 (\mu_r - \mu_{r-1})}{\tilde{w}_{r_2}},$$

donde $d_k = \frac{1}{2}\frac{1}{k-1}$ $\left(d_{k_{\text{máx}}} = \frac{1}{k_{\text{máx}}-1}\right)$ y $b_k = \frac{1}{2}\frac{1}{k}$ $\left(b_{k_{\text{mín}}} = \frac{1}{k_{\text{mín}}}\right)$ y siendo f_{BN} la densidad de una binomial negativa dada en (2.15) con parámetros $m = \nu_r + 1$ y $p = (m-1)/(\nu_r + m - 1.5)$ y $P_{\tilde{z}_i}$ la probabilidad dada en (2.24). Si en cambio, $\tilde{k} = k-1$ entonces, el valor de A_k^{sc} es la inversa del producto de estas expresiones teniendo en cuenta que, en este caso, cambian los valores candidatos.

En el paso 4 del algoritmo `RJHEr`, se genera, de nuevo, un valor candidato para el tamaño de la mixtura, $\tilde{k}$, eligiendo aleatoriamente entre una nacimiento o muerte de una componente de la mixtura. Si se produce

un nacimiento, se genera un peso, w_r^*, de una distribución beta, $Be(1,k)$, y los pesos restantes se reescalan de modo que sumen 1, es decir se multiplican por $(1-w_r^*)$. Los parámetros, ν_r^* y μ_r^*, de la nueva componente se generan de la distribución a priori,

$$w_r^* \sim Be(1,k), \qquad \nu_r^* \sim Geo(\vartheta), \qquad \mu_r^* \sim G(\alpha,\beta). \tag{2.26}$$

En caso de producirse una muerte, se elige al azar una componente vacía entre las existentes y se elimina. En este caso, los pesos se reescalan dividiendo por $(1-w_r^*)$, donde w_r^* es, ahora, el peso de la componente eliminada. La probabilidad de aceptar el movimiento es mín $\{1, A_k^{bd}\}$, donde A_k^{bd} es la expresión análoga para (2.25). Si, por ejemplo, se propone un nacimiento, es decir si $\tilde{k} = k+1$, se obtiene que,

$$\begin{aligned} A_k^{bd} &= \frac{f(k+1)}{f(k)} \times \frac{\Gamma(k+1)}{\Gamma(k)} \times (1-w_r^*)^n \times (k+1) \times \\ &\quad \times \frac{d_{k+1}/(k_0+1)}{b_k} \times \frac{1}{f_{Be}} \times (1-w_r^*)^{k-1}, \end{aligned}$$

donde k_0 es el número de componentes vacías antes de producirse el nacimiento y f_{Be} es la función de densidad de una $Be(1,k)$, por tanto, se obtiene que,

$$A_k^{bd} = \frac{f(k+1)}{f(k)} \times (1-w_r^*)^n \times (k+1) \times \frac{d_{k+1}/(k_0+1)}{b_k},$$

donde $f(k)$ es la distribución a priori de k. Si, por el contrario, se propone la muerte de una componente vacía, es decir, si $\tilde{k} = k-1$ entonces, el valor de A_k^{bd} es la inversa de esta expresiones teniendo en cuenta que, en esta situación, w_r^* es el peso de la componente que va a desaparecer y (k_0+1) es el número de componentes vacías antes de la muerte.

2.4.1.2. Método RJMCMC para ajustar una distribución MGE.

A continuación, se describe un algoritmo basado en el método del salto reversible para inferir sobre el número de componentes de la distribución MGE, véase (2.3). La estructura es similar a la del algoritmo `RJHEr`, aunque existen diferencias substanciales originadas por la ampliación del conjunto de datos faltantes, $(\mathbf{z},\mathbf{y})$. El paso 3, que incluye el movimiento de separación y combinación de componentes y el paso 4, que corresponde al movimiento de nacimiento y muerte de componentes vacías, se describen a continuación. Nótese que en estos pasos el valor del hiperparámetro permanece constante. El algoritmo se ilustra gráficamente en la Figura 2.5.

`ALGORITMO RJMGE.`

1. `Fijar valores iniciales para` $L^{(0)}, \mathbf{P}^{(0)}, \boldsymbol{\mu}^{(0)}$.
2. `ALGORITMO MGE: Generar valores de` $\left(\mathbf{z}^{(j+1)}, \mathbf{y}^{(j+1)}, \mathbf{P}^{(j+1)}, \boldsymbol{\mu}^{(j+1)}\right)$ `para` $L^{(j)}$ `fijo.`
3. `Separar una componente de la mixtura en dos, o combinar dos en una.`
4. `Nacimiento o muerte de una componente vacía.`
5. $j = j+1$. `Ir a 2.`

En el paso 3, si se propone un movimiento de combinación, se escogen aleatoriamente dos fases consecutivas (r_1, r_2) de la sucesión de estados que constituye el modelo MGE, véase (2.2), y se combinan en una sola modificando los parámetros del modo siguiente,

1. $\tilde{P} = P_{r_1} + P_{r_2}$.

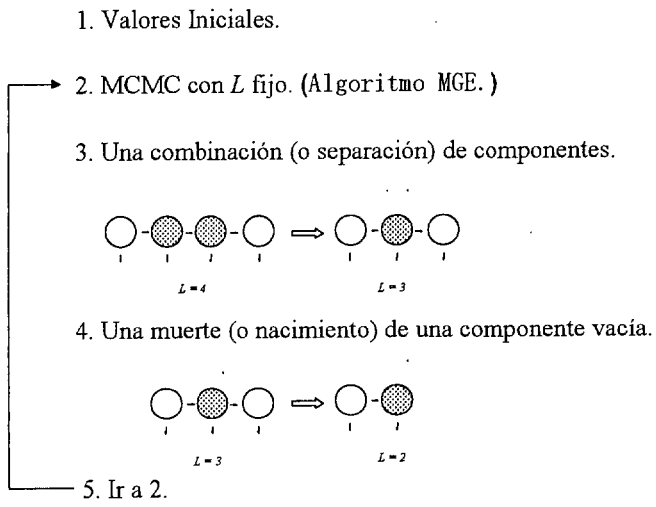


Figura 2.5: Algoritmo **RJMGE** asociado al modelo de mixtura MGE y basado en las técnicas de salto reversible. Las fases en blanco representan componentes vacías y en gris, fases con datos asociados

2. $\frac{1}{\tilde{\mu}} = \frac{1}{\mu_{r_1}} + \frac{1-\sum_{s=1}^{r_1} P_s}{1-\sum_{s=1}^{r_1-1} P_s} \frac{1}{\mu_{r_2}}$.

Se escogen estos movimientos para preservar los momentos de orden 0 y 1 de la variable de interés, X. Obsérvese que en el segundo paso se obtiene que $\tilde{\mu}_r \leq \mu_{r_1}$, esta restricción tiene que ser incluida en el movimiento inverso, que es el de separación, para mantener la reversibilidad. Por otra parte, en el movimiento de combinación, las variables latentes $(\mathbf{z}, \mathbf{y})$ se modifican de manera natural. Es decir, por una parte, todos los datos que abandonaban la sucesión de fases en alguno de los estados, $z_i = r_1$ ó r_2, abandonan, ahora, en el nuevo estado resultado de la combinación, esto es, $\tilde{z}_i = r$. Y, por otra parte, el tiempo, $\tilde{y}_{ir}$, que permanece la observación i-ésima en el nuevo estado será la suma de los tiempos en los estados que se pretenden combinar, esto es, $y_{ir_1} + y_{ir_2}$.

Si se propone un movimiento de separación, se escoge al azar una componente, r, para dividirse en dos. Con el fin de originar el valor de los nuevos parámetros, se generan dos valores, u_1 y u_2 de una distribución uniforme definida en el intervalo $(0,1)$ y se efectúan las siguientes modificaciones,

1. $\tilde{P}_{r_1} = u_1 P_r, \qquad \tilde{P}_{r_2} = (1-u_1) P_r.$

2. $\tilde{\mu}_{r_1} = \frac{\mu_r}{u_2}, \qquad \tilde{\mu}_{r_2} = \frac{\mu_r}{1-u_2} \frac{1-u_1 P_{r_1} - \sum_{s=1}^{r_1-1} P_s}{1-\sum_{s=1}^{r_1-1} P_s}.$

Obsérvese que, con estas transformaciones se obtiene, como se pretendía, que $\tilde{\mu}_{r_1} \geq \mu_r$, y, además, si se aplicase el movimiento de combinación sobre estos valores candidatos se obtendrían los valores originales. En cuanto a las transformaciones para las variables latentes, $(\mathbf{z}, \mathbf{y})$, se desarrollan análogamente al proceso de localización del muestreo de Gibbs descrito en la Sección 2.3.3. En concreto, si $z_i = r$, entonces, la observación i-ésima se asigna a cada una de las componentes, r_1 ó r_2, con probabilidad,

$$P(\tilde{z}_i = r_j) \propto \tilde{P}_{r_j} f_{r_j}(x_i \mid \tilde{\mu}_1, ..., \tilde{\mu}_r), \qquad j = 1, 2, \tag{2.27}$$

donde f_{r_j} es la función de densidad dada en (2.5). Por otro lado, si $\tilde{z}_i > r_1$, se propone dividir el tiempo, y_{ir}, de permanencia en el estado r de la observación i-ésima en dos tiempos, $(\tilde{y}_{ir_1}, \tilde{y}_{ir_2})$, en cada uno de los dos nuevos estados. Si se supone, sin pérdida de generalidad, que $\tilde{\mu}_{r_1} = \text{mín}\left\{\tilde{\mu}_{r_1}, \tilde{\mu}_{r2}\right\}$, entonces, análogamente

a (2.23), se tiene que la distribución de $\tilde{y}_{ir_2}$ es una exponencial truncada,

$$f(\tilde{y}_{ir_2}) \propto (\tilde{\mu}_{r_2} - \tilde{\mu}_{r_2}) \exp\{-(\tilde{\mu}_{r_2} - \tilde{\mu}_{r_1}) y_{ir}\} \times \mathbf{I}_{[\tilde{y}_{ir_2} < y_{ir}]}, \tag{2.28}$$

y, naturalmente, $\tilde{y}_{ir_1} = y_{ir} - \tilde{y}_{ir_2}$.

Obsérvese que el hiperparámetro, η, no sufre modificaciones ante los movimientos de combinación y separación. Finalmente, se acepta el movimiento de separación o combinación con probabilidad mín $\{1, A_L^{sc}\}$, donde,

$$A_L^{sc} = \frac{f(\bar{\theta} \mid \mathbf{x}) p(\bar{\theta}, \theta)}{f(\theta \mid \mathbf{x}) p(\theta, \bar{\theta})},$$

y donde, equivalentemente a (2.16), $f(\theta \mid \mathbf{x})$ es la distribución a posteriori de los parámetros y $p(\theta, \bar{\theta})$ es la probabilidad de moverse desde $\theta = (L, \mathbf{P}, \boldsymbol{\mu})$ a $\bar{\theta} = (\tilde{L}, \tilde{\mathbf{P}}, \tilde{\boldsymbol{\mu}})$. Si, en particular, $\tilde{L} = L + 1$, se obtiene que,

$$A_L^{sc} = \frac{f(L+1)}{f(L)} \times \frac{\Gamma(L+1)}{\Gamma(L)} \times \frac{e^{-\eta\tilde{\mu}_{r_1}} e^{-\eta\tilde{\mu}_{r_2}}}{e^{-\eta\mu_r}} \times \frac{\tilde{P}_{r_1}^{\tilde{n}_{r_1}} \tilde{P}_{r_2}^{\tilde{n}_{r_2}}}{P_r^{n_r}} \times \frac{\tilde{\mu}_{r_1}^{\tilde{n}'_{r_1}} \exp\{-\tilde{n}'_{r_1}\tilde{\mu}_{r_1}\tilde{y}_{ir_1}\} \tilde{\mu}_{r_1}^{\tilde{n}'_{r_1}} \exp\{-\tilde{n}'_{r_2}\tilde{\mu}_{r_2}\tilde{y}_{ir_2}\}}{\mu_r^{n'_r} \exp\{-n'_r\mu_r y_{ir}\}} \times$$

$$\times \frac{d_{L+1}}{b_k \prod_{i:z_i=r} P_{\tilde{z}_i} \prod_{i:\tilde{z}_i>r_1} f_{\tilde{\mathbf{y}}_i}} \times \frac{P_r\mu_r}{u_2^2(1-u_2)^2} \frac{1 - \sum_{s=1}^{r-1} P_s - P_r u_1}{1 - \sum_{s=1}^{r} P_s}$$

donde $d_L = \frac{1}{2}\frac{1}{L-1}$ ($d_{L_{\text{máx}}} = \frac{1}{L_{\text{máx}}-1}$) y $b_L = \frac{1}{2}\frac{1}{L}$ ($b_{L_{\text{mín}}} = \frac{1}{L_{\text{mín}}}$) y donde $P_{\tilde{z}_i}$ y $f_{\tilde{\mathbf{y}}_i}$ son las densidades asociadas a la asignación de los datos faltantes a las observaciones afectadas por el movimiento dadas por (2.27) y (2.28), respectivamente.

El paso 4, correspondiente al movimiento de nacimiento y muerte de componentes vacías es equivalente al movimiento análogo del algoritmo RJHEr. Sin embargo, en este caso, las componentes vacías se asocian a fases de la distribución MGE que no son atravesadas por ninguna observación y por tanto, las fases vacías se encuentran, necesariamente, al final de la cadena de estados que constituye el modelo MGE. Si se produce un nacimiento de una componente vacía, se genera un estado nuevo, que se sitúa en última posición, con parámetros (P^*, μ^*) tales que,

$$P_r^* \sim Be(1, L), \qquad \mu_r^* \sim Exp(\eta).$$

Una muerte origina la desaparición de la última fase si se trata de una fase vacía, es decir, si ninguna observación ha permanecido tiempo en ella. Tanto en el movimiento de nacimiento como en el de muerte de una fase vacía, los pesos de las fases restantes se reescalan de modo que sumen 1. En este algoritmo, la probabilidad de aceptar un nacimiento de una componente vacía es igual a mín $\{1, A_L^{bd}\}$, donde,

$$A_L^{bd} = \frac{f(L+1)}{f(L)} \times (1 - P_r^*)^n \times \frac{d_{L+1}}{b_L}.$$

2.4.2. Métodos en tiempo continuo (BDMCMC).

En esta Subsección, se proponen dos procedimientos alternativos a los algoritmos RJHEr y RJMGE, descritos en la Sección anterior. La finalidad es la misma, construir una cadena de Markov que se mueva por la distribución a posteriori del número de términos en la mixtura. Sin embargo, en esta Sección, se hace uso de una técnica alternativa a los métodos de salto reversible que son los denominados métodos BDMCMC que se proponen en Stephens (2000a). El objetivo posterior es establecer comparaciones entre los algoritmos RJMCMC y BDMCMC aplicados a los modelos de distribución HEr y MGE de manera que se puedan determinar diferencias entre ambas metodologías tanto en la implementación de los algoritmos como en los resultados obtenidos sobre el movimiento y convergencia de la cadena, y por supuesto, en la estimación de la densidad correspondiente.

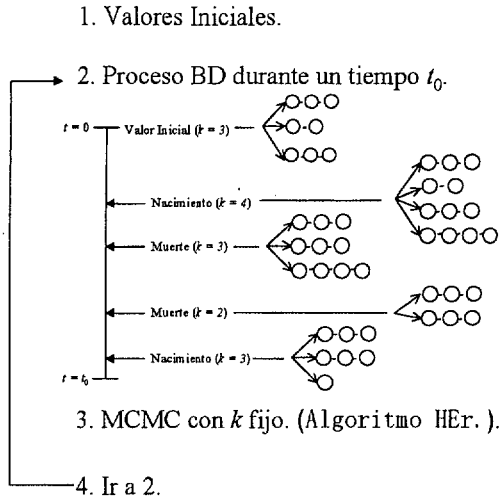


Figura 2.6: Algoritmo BDHEr asociado al modelo de mixtura HEr y de tipo BDMCMC.

Con el fin de permitir movimientos en la dimensión paramétrica, en los métodos BDMCMC, se sustituyen los movimientos del tipo Metropolis Hastings propios del RJMCMC, por un proceso de nacimiento y muerte cuyo espacio de estados es el espacio paramétrico, incluida la dimensión de la mixtura. Además, la distribución estacionaria de este proceso debe coincidir con la distribución conjunta a posteriori de los parámetros. De este modo, la inclusión de etapas MCMC con dimensión paramétrica fija no es estrictamente necesaria aunque es recomendable para mejorar el movimiento de la cadena. Los movimientos en el número de términos de la mixtura se producen mediante nacimientos y muertes de las componentes de la misma, los cuales acontecen en tiempo continuo. Por tanto, la estrategia de aceptar y rechazar candidatos para el número de términos en la mixtura, característica de los métodos RJMCMC, se reemplaza por la longitud del periodo de tiempo (*holding time*) en cada estado paramétrico.

2.4.2.1. Método BDMCMC para ajustar una distribución HEr.

Se pretende construir una cadena de Markov cuya distribución estacionaria sea la distribución a posteriori de los parámetros, $(k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu})$, de una mixtura de distribuciones Erlang, véase (2.1). Para ello, se puede combinar un proceso de nacimiento y muerte (BD), cuya distribución estacionaria sea $f(k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu} \mid \mathbf{x})$, con un método MCMC en el que el valor de k permanezca constante. Este es el procedimiento que se propone en Stephens (2000a) para mixturas de normales. El esquema del algoritmo es el siguiente y el proceso BD adaptado para mixturas de distribuciones Erlang se describe a continuación. Este algoritmo se representa de forma simbólica en la Figura 2.6.

ALGORITMO BDHEr.

1. Fijar valores iniciales para $k^{(0)}, \mathbf{w}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\nu}^{(0)}$.
2. Simular un proceso BD durante un tiempo fijo t_0 partiendo de $\mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\nu}^{(j)}$.
3. Fijar $k^{(j+1)}$ igual a la dimensión al terminar el periodo t_0 del proceso BD.
4. ALGORITMO HEr: Generar valores de $\left(\mathbf{z}^{(j+1)}, \mathbf{w}^{(j+1)}, \boldsymbol{\mu}^{(j+1)}, \boldsymbol{\nu}^{(t+1)}\right)$ para $k^{(j+1)}$ fijo.
5. $j = j + 1$. Ir a 2.

El paso 2 es el proceso de nacimiento y muerte, que se simula durante un periodo de tiempo fijo que se puede establecer, por ejemplo, $t_0 = 1$. En el transcurso de este periodo se producen nacimientos y muertes de las componentes de la mixtura en tiempo continuo, que permiten variar la dimensión, k, de los parámetros.

Los nacimientos se producen con una tasa constante, γ, que se puede fijar, por ejemplo, igual a 2. Un nacimiento incrementa el número de componentes de la mixtura en una unidad. Se genera el peso de la nueva componente de una distribución beta con parámetros $(1, k)$ y el resto de los parámetros se generan de la distribución a priori de la misma manera que se generaban los parámetros de una nueva componente en el movimiento de nacimiento de componentes vacías del algoritmo `RJHEr`, véase (2.26). Obsérvese que los nacimientos constituyen un proceso de Poisson de tasa γ, y por tanto, reproducir el proceso un número determinado de veces es equivalente a multiplicar la tasa de nacimiento por ese número. Por esta razón, al igual que Stephens (2000a), se ha fijado $t_0 = 1$. Como es de esperar, en la práctica se obtiene que, al aumentar el valor de la tasa de nacimiento, γ, la cadena se mueve mejor entre los valores posibles para L, sin embargo, este aumento de γ da lugar a un incremento del tiempo de computación.

Las muertes se producen con una tasa, δ, que varía a lo largo del proceso BD. Una muerte reduce el número de componentes de la mixtura en una unidad. Esta tasa, δ, se obtiene como resultado de la suma de las tasa individuales de cada término, $\delta = \sum_{r=1}^{k} \delta_r$. La tasa de muerte de una componente determinada, r_0, viene dada por,

$$\delta_{r_0} = \gamma \frac{p(k-1)}{kp(k)} \prod_{i=1}^{n} \left(\frac{\sum_{\substack{r=1 \\ r \neq r_0}}^{k} \frac{w_r}{1-w_{r_0}} Er(x_i \mid \nu_r, \mu_r)}{\sum_{r=1}^{k} w_r Er(x_i \mid \nu_r, \mu_r)} \right), \qquad \text{para } r_0 = 1, ..., k.$$

Obsérvese que si se escoge un distribución de Poisson truncada a priori para k con tasa igual a la tasa de nacimiento, γ, es decir,

$$p(k) \propto \frac{\gamma^k}{k!}, \qquad \text{para } k = 1, ..., k_{\text{máx}},$$

entonces la tasa de muerte, δ_{r_0}, será igual a la razón de verosimilitudes de la distribución HEr con y sin la componente r_0. En general, la tasa de muerte de una componente determinada será pequeña si ésta explica en gran medida el comportamiento de los datos y será grande en caso contrario. De este modo, las componentes que no expliquen datos morirán rápidamente y viceversa.

Los procesos de nacimiento y muerte son procesos de Poisson independientes y por tanto, el tiempo que transcurre hasta el próximo nacimiento ó muerte está distribuido exponencialmente con media $1/(\delta + \gamma)$ y se produce un nacimiento o una muerte con probabilidades proporcionales a γ y δ, respectivamente. Por tanto, es fácil simular un proceso con las características que se acaban de describir y el procedimiento se resume esquemáticamente a continuación.

2. Proceso BD simulado durante t_0.
   a. Comenzar en $\left(\mathbf{w}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\nu}^{(t)}\right)$.
   b. Calcular las tasa de muerte para $r = 1, ..., k^{(j)}$.
   c. Simular un tiempo $Exp(\delta + \gamma)$ hasta el próximo nacimiento o muerte.
   d. Generar si se trata de nacimiento $(prob = \gamma/(\gamma + \delta))$ o muerte $(prob = \delta/(\gamma + \delta))$.
   e. Modificar $\mathbf{w}$, $\boldsymbol{\mu}$ y $\boldsymbol{\nu}$ para reflejar el nacimiento o muerte.
   f. Si el tiempo de ejecución es menor que t_0 ir a (b).

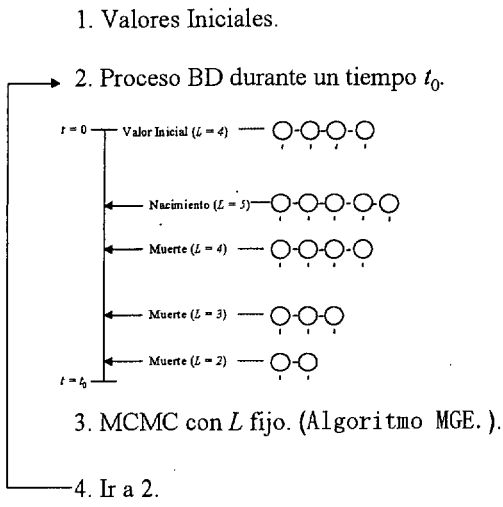


Figura 2.7: Algoritmo BDMGE asociado al modelo de mixtura MGE y de tipo BDMCMC.

2.4.2.2. Método BDMCMC para ajustar una distribución MGE.

Se describe, ahora, de modo abreviado, un algoritmo de tipo BDMCM para obtener una muestra de la distribución a posteriori de los parámetros $(L, \mathbf{P}, \boldsymbol{\mu})$ del modelo MGE, véase (2.3). Para la distribución Coxiana es necesario modificar el enfoque del algoritmo puesto que, a pesar de ser un modelo de mixtura, no resulta natural el nacimiento y muerte de componentes sino de estados o fases exponenciales de la cadena que constituye el modelo MGE, véase (2.2). Además, es necesario incorporar los movimientos del hiperparámetro, η, que, como se propone en Stephens (2000a), permanecerá fijo durante el proceso BD. La estructura es muy similar a la del algoritmo BDHEr. El algoritmo se ilustra en la Figura 2.7.

ALGORITMO BDMGE.

1. Fijar valores iniciales para $L^{(0)}, \mathbf{P}^{(0)}, \boldsymbol{\mu}^{(0)}$.
2. Simular un proceso BD durante un tiempo fijo t_0 partiendo de $L^{(j)}, \mathbf{P}^{(j)}, \boldsymbol{\mu}^{(j)}$ y en el que $\eta^{(j)}$ permanece constante.
3. Fijar $L^{(j+1)}$ igual a la dimensión al terminar el periodo t_0 del proceso BD.
4. ALGORITMO MGE: Generar valores de $\left(\mathbf{z}^{(j+1)}, \mathbf{y}^{(j+1)}, \mathbf{P}^{(j+1)}, \boldsymbol{\mu}^{(j+1)}, \eta^{(j+1)}\right)$ para $L^{(j+1)}$ fijo.
5. $j = j + 1$. Ir a 2.

En el proceso BD, simulado en el paso 2, se producen nacimientos y muertes de los estados o fases de la distribución MGE. Si se produce un nacimiento, se escoge al azar la posición del nuevo estado entre todas las posibles. Obsérvese que el nacimiento de una nueva fase origina el nacimiento de una nueva componente en la mixtura. Se genera el peso de la nueva componente de una distribución beta con parámetros $(1, L)$ y se genera la tasa del nuevo estado a partir de una exponencial de tasa igual al hiperparámetro η, que es su distribución a priori.

Una muerte reduce el número de componentes en la mixtura en una unidad originada por la desaparición

de uno de los estados de la distribución MGE. La tasa de muerte de cada estado viene dada por,

$$\delta_{r_0} = \gamma \frac{p(L-1)}{Lp(L)} \prod_{i=1}^{n} \left(\frac{\sum_{\substack{r=1 \\ r \neq r_0}}^{L} \frac{P_r}{1-P_{r_0}} f_{r_-}(x_i \mid \boldsymbol{\mu})}{\sum_{r=1}^{L} P_r f_r(x_i \mid \boldsymbol{\mu})} \right), \qquad \text{para } r_0 = 1, ..., L,$$

donde f_r es la función de densidad de la suma de r exponenciales, $\sum_{r=1}^{L} Y_r$, dada en (2.5), y f_{r_-} representa la densidad de la suma de estas r exponenciales sin la fase r_0, es decir, la densidad de $\sum_{\substack{r=1 \\ r \neq r_0}}^{L} Y_r$. Al igual que antes, las tasas de muerte de cada fase serán pequeñas si el tiempo exponencial que representan explica el valor de los datos y viceversa.

2.4.3. Estimación a partir de los algoritmos con dimensión paramétrica variable.

En esta Subsección, se describe, brevemente, cómo hacer inferencia a partir de las muestras MCMC obtenidas mediante los algoritmos de las Subsecciones anteriores.

Si el punto de partida es una realización de una cadena MCMC construida para ajustar un conjunto de n observaciones, $\mathbf{x} = \{x_1, ..., x_n\}$, a una mixtura de distribuciones Erlang obtenida a partir de los algoritmos `RJHEr` ó `BDHEr`, la distribución predictiva de la variable de interés, X, se puede estimar mediante la siguiente expresión,

$$f(x \mid \mathbf{x}) = \frac{1}{J} \sum_{j=1+B}^{J+B} \sum_{r=1}^{k^{(j)}} w_r^{(j)} Er(x \mid \nu_r^{(j)}, \mu_r^{(j)}) \tag{2.29}$$

donde J es el tamaño de la muestra MCMC en equilibrio. Obsérvese que esta expresión es similar a la dada en (2.18) pero, en este caso, la dimensión de los parámetros de cada observación, $\left(k^{(j)}, \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\nu}^{(j)}\right)$, no es constante.

Se puede desarrollar también inferencia para el tamaño de la mixtura, k, estimando la distribución marginal a posteriori mediante,

$$P(k \mid \mathbf{x}) \approx \frac{1}{J} \# \left\{ j : k^{(j)} = k \right\}. \tag{2.30}$$

Por ejemplo, la probabilidad a posteriori de tratarse de una simple distribución Erlang sería $P(k = 1 \mid \mathbf{x})$. Si esta probabilidad es lo suficientemente grande, se podría considerar que la distribución Erlang es el modelo verdadero de los datos. Análogamente, la probabilidad a posteriori de que el conjunto de datos provenga de una exponencial se puede aproximar mediante,

$$P(\nu = 1, k = 1 \mid \mathbf{x}) \approx \frac{1}{J} \# \left\{ j : k^{(j)} = 1 \text{ y } \nu^{(j)} = 1 \right\}, \tag{2.31}$$

y, en particular, si se asume que una componente es suficiente para explicar los datos, la probabilidad de que la distribución sea exponencial se puede estimar con,

$$P(\nu = 1 \mid \mathbf{x}, k = 1) \approx \frac{1}{J_1} \# \left\{ \nu^{(j)} = 1 \right\}, \tag{2.32}$$

donde $J_1 = \# \left\{ j : k^{(j)} = 1 \right\}$ es el número de iteraciones con dimensión paramétrica igual a 1. Del mismo modo que antes, si la probabilidad obtenida en (2.31) es muy alta, se puede asumir que los datos son exponenciales.

En estos casos, si el objetivo es hacer predicción Bayesiana en sistemas de colas, es posible desarrollar técnicas más sencillas para sistemas más simples como son el sistema M/Er/1, cuyas características son

relativamente fáciles de estimar, véase, por ejemplo, Ríos et al. (1998) o el sistema Er/M/1, véase Wiper (1998), o el sistema M/M/1, véase Armero y Bayarri (1994a), dependiendo del caso. En estos sistemas, condicionando en el valor de los parámetros, se conocen algunos resultados exactos para el tamaño del sistema en equilibrio, el tiempo de espera y la longitud del periodo de ocupación y en las referencias anteriores se muestra cómo estimar estas medidas con un enfoque Bayesiano.

Análogamente, si se tiene una muestra MCMC de los valores de los parámetros de una distribución MGE obtenida a partir de los algoritmos `RJMGE` y `BDMGE`, se puede estimar la distribución predictiva de X mediante,

$$f(x \mid \mathbf{x}) = \frac{1}{J} \sum_{j=B+1}^{J+B} \sum_{r=1}^{L^{(j)}} P_r^{(j)} f_r \left(x \mid \mu_1^{(j)}, ..., \mu_r^{(j)} \right), \tag{2.33}$$

donde f_r es la densidad dada en (2.5). Y, de la misma manera, se pueden estimar las probabilidades a posteriori del número de fases L con,

$$P(L \mid \mathbf{x}) \approx \frac{1}{J} \# \left\{ j : L^{(j)} = L \right\}. \tag{2.34}$$

En este caso, la probabilidad de estar considerando un conjunto de datos exponenciales es $P(L = 1 \mid \mathbf{x})$, y la probabilidad de que provengan de una distribución Erlang sería $P(\mu_1 = \ldots = \mu_L \mid \mathbf{x})$, pero como la distribución conjunta de las tasas, $\boldsymbol{\mu}$, es continua se tiene que esta probabilidad es igual a cero. Sin embargo, se puede considerar que los datos proceden de una distribución Erlang si se obtiene un valor alto para la probabilidad conjunta a posteriori de que la diferencia de todos los pares de tasas en valor absoluto sea menor que una cota fijada previamente, c_0,

$$P(|\mu_s - \mu_r| < c_0,\ \forall r, s \mid \mathbf{x}) \approx \frac{1}{J} \# \left\{ j : \left| \mu_s^{(j)} - \mu_r^{(j)} \right| < c_0, \text{ para } r, s = 1, ..., L^{(j)} \right\}.$$

2.5. Ilustración numérica.

En esta Sección, se ilustra el funcionamiento de los distintos procedimientos descritos en este Capítulo para la estimación Bayesiana de densidades continuas y positivas. Para ello, se simulan cinco conjuntos de datos procedentes de 5 modelos distintos de distribuciones y con las características que se describen a continuación. También, se considera una muestra de observaciones reales que describen los tiempos de ocupación en un hospital geriátrico.

1. 100 datos generados de una distribución exponencial de media 1.0.
2. 100 observaciones generadas de una mixtura de distribuciones Erlang con, $\mathbf{w} = (0.3, 0.35, 0.35)$, $\boldsymbol{\mu} = (1/2, 1/6, 1/12)$ y $\boldsymbol{\nu} = (10, 15, 25)$.
3. 100 observaciones generadas de una distribución Coxiana con parámetros $\mathbf{P} = (0.1, 0.9)$ y $\boldsymbol{\mu} = (5, 30)$.
4. 100 datos iguales a 1.0 procedentes de una distribución degenerada.
5. 100 datos de una distribución Weibull, $Weib(a, b)$, con parámetros $a = 1.5$ y $b = 1.5$, véase (2.35).
6. 1092 datos de la duración (en días) de la estancia de enfermos en un hospital geriátrico.

Obsérvese que el caso 1 es la forma más sencilla de los modelos de distribución HEr y MGE. El caso 2 es una distribución HEr que podría expresarse como una distribución MGE, aunque con 50 fases. El caso 3 es una distribución MGE que no pertenece a la familia de distribuciones HEr. La distribución degenerada, en el caso 4, se puede considerar como el límite de una distribución Erlang $Er(\nu, 1)$ donde $\nu \to \infty$ y, por tanto,

también es el límite de una distribución Coxiana con $P_L = 1$, con todas las tasas iguales y con $L \to \infty$. El caso 5 corresponde a una distribución Weibull, $Weib(a, b)$, cuya densidad viene dada por,

$$f(x) = abx^{b-1} \exp\left\{-ax^b\right\}, \qquad x > 0. \tag{2.35}$$

Esta distribución no pertenece a ninguna de las dos familias de mixturas que se han propuesto, pero es un ejemplo de distribución que se utiliza frecuentemente para ajustar datos con colas pesadas. Los datos reales del caso 6 corresponden al Hospital St. George de Londres en el transcurso de 1965 hasta 1984. Esta muestra se describe detalladamente en el Capítulo 4 y es un subconjunto de los datos analizados en Taylor et al. (2000) que se pueden obtener en la siguiente dirección de Internet, `http://www.blackwellpublishers.co.uk/rss/`.

Se pretende ajustar los modelos de distribución HEr y MGE a cada conjunto de datos utilizando los algoritmos que se han propuesto. En cada caso, se supone una distribución uniforme discreta a priori para el número de componentes en la mixtura con un valor máximo para el mismo igual a 10. Para cada conjunto de datos, se ejecutan, en primer lugar, los algoritmos `RJHEr` y `RJMGE`, que son de tipo RJMCMC, con el fin de comparar las estimaciones obtenidas con los modelos de distribución HEr y MGE. A continuación, se ejecutan también los algoritmos `BDHEr` y `BDMGE` con el objetivo de comparar el funcionamiento de estos métodos, que son de tipo BDMCMC, con los de tipo RJMCMC. Se fija en cada cadena 100000 iteraciones para obtener convergencia (*burn-in*), y 100000 iteraciones "en equilibrio". Los algoritmos `RJHEr` y `RJMGE`, programados en FORTRAN, requieren menos de 20 minutos para su realización en un Pentium IV, para los 5 casos simulados y algo más de media hora para el conjunto de datos reales. El tiempo de ejecución se incrementa en algo más de cinco minutos, aproximadamente, para los algoritmos `BDHEr` y `BDMGE`.

2.5.1. Ajuste de distribuciones HEr y MGE.

Las Figuras 2.8 y 2.9 ilustran, con dos tipos de líneas discontinuas, las estimaciones de las distribuciones predictivas asumiendo una distribución HEr, véase (2.29), y suponiendo un modelo de distribución MGE, véase (2.33). También, se muestran con línea continua las verdaderas densidades en los casos simulados y no degenerados. Se puede apreciar que la estimación de la densidad de los datos exponenciales, en el caso 1, es muy similar a la verdadera función utilizando cualquiera de los dos modelos de distribución. En el caso 2, se observa que la estimación que se ajusta adecuadamente a la verdadera función de densidad corresponde a la mixtura de distribuciones Erlang, que es el verdadero modelo generador de los datos. Sin embargo, nótese que la distribución Coxiana tiene problemas para captar la multimodalidad. Hay que tener en cuenta que, como se comentó en la Sección 2.2.3, para que este modelo aproxime bien este tipo de comportamiento se requiere un número muy elevado de fases, L, y, en estos casos, se ha supuesto a priori que el soporte de L se limita al intervalo $[1, 10]$. En el caso 3, se aprecia que ambos modelos de distribución, HEr y MGE, describen favorablemente el comportamiento de las observaciones. No obstante, la distribución MGE da lugar a un modelo más sencillo que la distribución HEr ya que la probabilidad a posteriori para el tamaño de la mixtura, k y L, se concentra, para los dos modelos de distribución, entre los valores 2, 3 y 4, lo cual implica que el número de parámetros necesarios para una buena aproximación utilizando la distribución HEr es superior a los requeridos cuando se hace uso del modelo MGE. En el caso degenerado, como era de esperar, las densidades estimadas están concentradas alrededor del valor $x = 1.0$, sin embargo, con la distribución HEr se obtiene un ajuste mejor puesto que este modelo de distribución permite una varianza muy pequeña mediante valores elevados del parámetro entero ν, mientras que la distribución MGE necesita, para conseguir lo mismo, un número muy grande para L, como se ha comentado anteriormente. Los datos procedentes de una distribución Weibull, en el caso 5, se ajustan favorablemente con cualquiera de los dos modelos de mixtura. Finalmente, en el caso 6, las dos estimaciones de los datos reales parecen captar adecuadamente la forma de la distribución para las observaciones superiores a 8. Sin embargo, se puede apreciar un sobreajuste de la distribución estimada con el modelo HEr para los valores inferiores a 8, asignando una componente para cada valor entero de la variable. Hay que tener en cuenta que el histograma muestra únicamente las observaciones menores de 150 días en el hospital, pero el conjunto total de datos es muy asimétrico a la

derecha con un valor máximo de 2658 días, como se muestra en la Figura 2.12 y en el Capítulo 4. Más comentarios sobre el origen de este sobreajuste se incluyen en la siguiente Subsección dentro del apartado destinado a la sensibilidad. La distribución MGE, en este caso, no tiene problemas para captar la forma de la distribución. Esta estimación se puede examinar más minuciosamente en la Figura 2.13.

En resumen, ambos modelos de mixturas, HEr y MGE, son, en general, adecuados para la descripción de datos con las características que se han citado. No obstante, a la hora de seleccionar entre las dos familias de distribuciones, hay que tener en cuenta que si los datos presentan multimodalidad puede haber dificultades para que modelo MGE sea totalmente favorable. Sin embargo, en muchos casos, la distribución MGE ofrece ventajas respecto a la distribución HEr en cuanto a simplicidad del modelo.

2.5.2. Comparación numérica de algoritmos.

Se pretende, ahora, hacer una comparación numérica de los algoritmos que se han propuesto. En concreto, el objetivo es examinar las diferencias entre los algoritmos de tipo RJMCMC (`RJHEr` y `RJMGE`) y BDMCMC (`BDHEr` y `BDMGE`). Recientemente, en Cappé et al. (2003), se han encontrado equivalencias importantes entre ambas metodologías. Para ello, se han englobado los métodos BDMCMC dentro de lo que se han denominado métodos MCMC en tiempo continuo (CTMCMC) y se ha demostrado que para cada algoritmo de tipo BDMCMC se puede construir una sucesión de algoritmos RJMCMC en tiempo discreto que convergen a él. Esta afirmación no implica, en la práctica, que ambos tipos de algoritmos se comporten de la misma manera.

Ciertamente, los algoritmos de tipo BDMCMC, han resultado más sencillos de implementar que los de tipo RJMCMC, puesto que se evitan cálculos de Jacobianos de transformaciones y la búsqueda de una elección razonable para el movimiento de separación y combinación de componentes. Sin embargo, hay que tener en cuenta, como se afirma en Cappé et al. (2003), que la simplicidad de la programación tiene su origen fundamentalmente en el tipo de movimiento de los algoritmos BDMCMC, que son siempre de nacimiento y muerte de componentes, y no en el diseño en tiempo continuo del algoritmo. Y, probablemente, si se hubiesen incluido movimientos de combinación y división en los algoritmos de tipo BDMCMC la complejidad hubiera sido similar. Por otro lado, se ha apreciado que los tiempos de computación son ligeramente superiores para los algoritmos de tipo BDMCMC. El motivo es, fundamentalmente, que en cada iteración es necesario calcular las tasas de muerte de cada componente y consecuentemente, evaluar tantas razones de verosimilitud como número de componentes. Concretamente, como se comentó anteriormente, en los ejemplos realizados, el tiempo de computación se incrementa aproximadamente entre cinco y diez minutos, para cada caso.

A pesar de estas diferencias, realmente, lo más interesante es examinar la convergencia y exploración del espacio de estados paramétrico de ambos algoritmos y, evidentemente, los resultados de las estimaciones. La Figura 2.10 compara las funciones de densidad estimadas para los ejemplos 2 y 3 utilizando los dos tipos de algoritmos y considerando en cada ejemplo el verdadero modelo generador de los datos. Como se puede observar ambas estimaciones son muy parecidas, casi indistinguibles. Un modo de evaluar la rapidez de movimiento en los algoritmos de tipo Metropolis, como son los métodos de salto reversible, es la tasa de valores candidatos aceptados. Si la tasa es muy alta, cada valor estará muy correlado con la iteración anterior y la cadena se moverá despacio. Las tasas de movimientos de separación o combinación aceptados en los algoritmos de salto reversible son 13.49 % para el ejemplo 2, que corresponde a la distribución HEr, y de 10.50 % para los datos del ejemplo 3, generados según un modelo MGE. Para movimientos en los que se cambia la dimensión de los parámetros estos valores resultan adecuados y reflejan que los valores candidatos que se proponen son razonables, véase Richardson y Green (1997). Estas tasas se pueden comparar, al igual que se hace en Stephens (2000a), con la proporción de iteraciones que cambian el valor de la dimensión de la distribución en un algoritmo BDMCMC. En concreto, se ha obtenido que estas proporciones son 49.19 %, para la muestra de la distribución HEr del caso 2 y 62.07 % para el caso 3 correspondiente a la muestra de MGE. Aunque las proporciones son superiores en los algoritmos de tipo BDMCMC esto no tiene por qué implicar que el movimiento de la cadena sea mejor ya que muchos de los estados que se visitan pueden no tener un peso significativo en el sentido de que el periodo de tiempo (*holding time*) que se permanece en

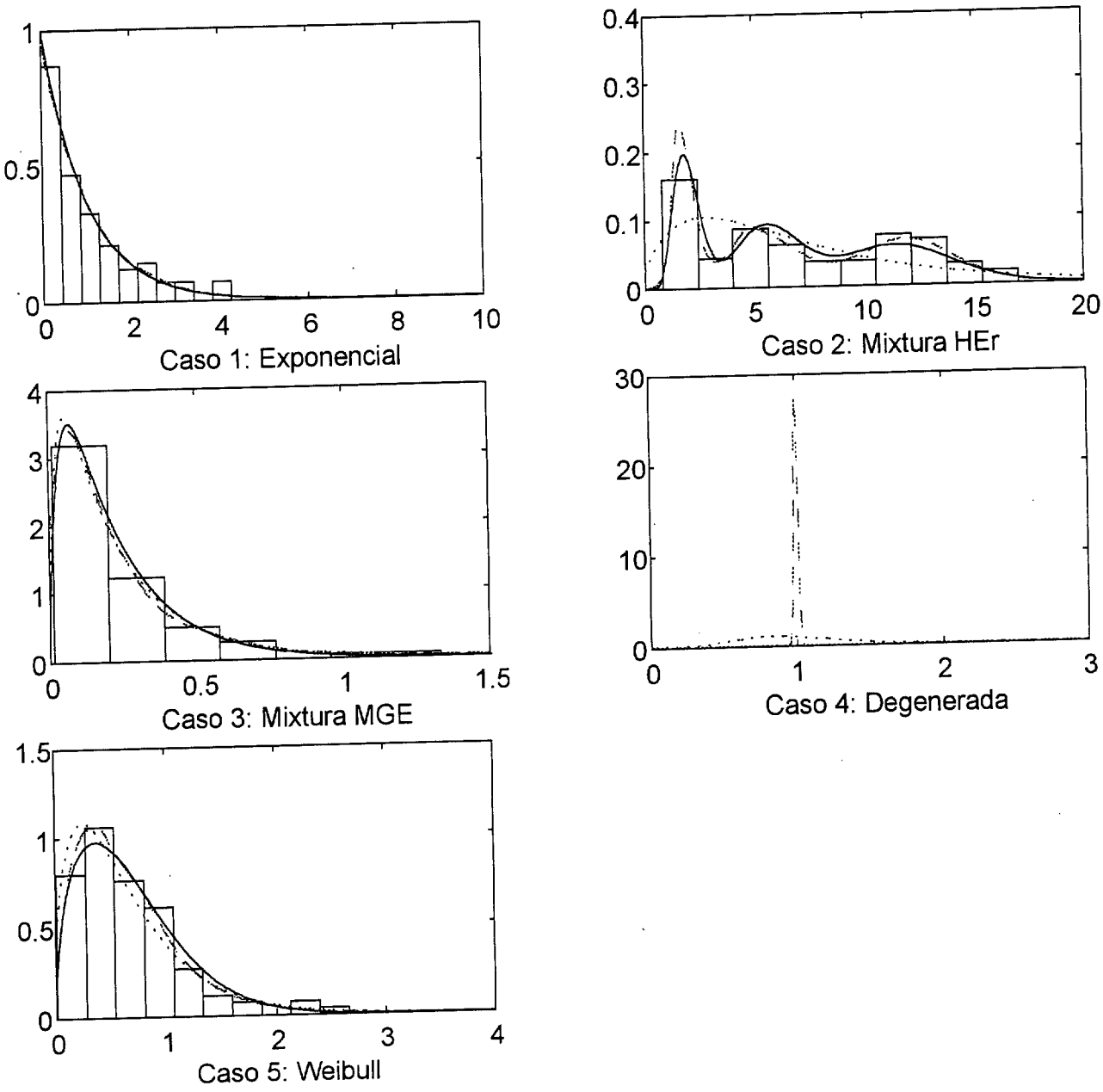


Figura 2.8: Histograma de los ejemplos simulados, funciones de densidad predictivas suponiendo los dos modelos de mixtura, HEr $(- -)$ y MGE $(\cdots)$, resultantes de los algoritmos `RJHEr` y `RJMGE`, respectivamente, y densidades verdaderas $(-)$ en los casos no degenerados.

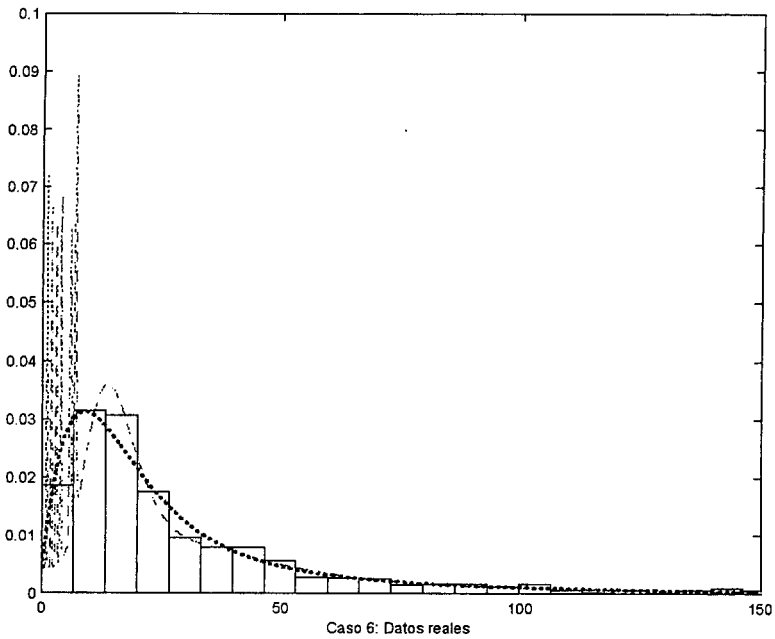


Figura 2.9: Histograma del ejemplo real considerado, funciones de densidad predictivas suponiendo los dos modelos de mixtura, HEr (− −) y MGE (· · ·), resultantes de los algoritmos `RJHEr` y `RJMGE`, respectivamente.

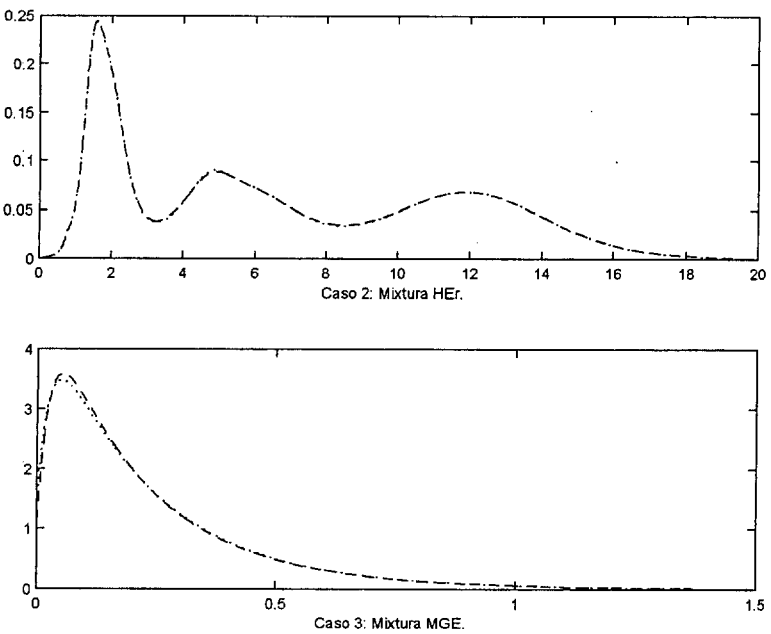


Figura 2.10: Comparación de las densidades estimadas utilizando los algoritmos de tipo RJMCMC (−−) y de tipo BDMCMC (· · ·) correspondientes a los algoritmos `RJHEr` y `BDHEr` (arriba) y `RJMGE` y `BDMGE` (abajo). Las estimaciones son muy parecidas, casi indistinguibles.

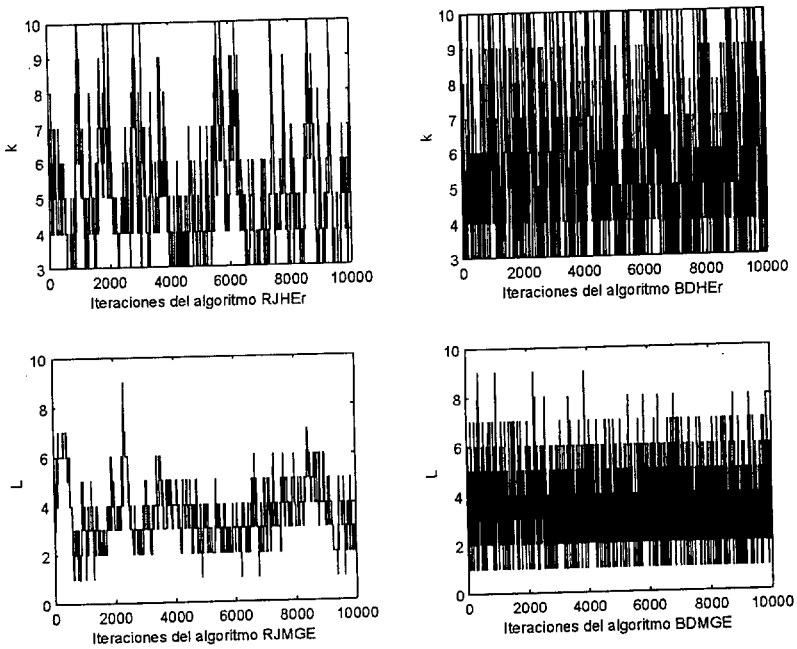


Figura 2.11: Cambios en el tamaño de la mixtura k (arriba) para los datos del caso 2 y cambios de L (abajo) para los datos del caso 3.

ellos puede ser muy breve, véase Cappé et al. (2003). En general, se ha comprobado que los algoritmos que suponen un modelo de mixtura HEr exploran mejor el espacio de estados que si se supone una distribución Coxiana, independientemente del tipo del algoritmo que se utilice. El motivo es que los movimientos en la dimensión de los parámetros en distribuciones MGE originan densidades más similares que los movimientos en la dimensión de distribuciones HEr, aunque la correlación entre los parámetros sea la misma.

La Figura 2.11 ilustra los valores de k y de L para las iteraciones en equilibrio de los algoritmos de tipo salto reversible (`RJHEr` y `RJMGE`) y los métodos en tiempo continuo (`BDHEr` y `BDMGE`) utilizados para las estimaciones de la Figura 2.10. Se puede apreciar que el valor de la dimensión paramétrica explora adecuadamente el espacio de estados visitando casi la totalidad del soporte de k y de L. Se observa que, aparentemente, los algoritmos de tipo BDMCMC exploran más rápidamente el espacio de estados. Como se ha comentado antes, no se puede concluir que la convergencia sea mejor para los algoritmos `BDHEr` y `BDMGE` puesto que muchos de los estados que se visitan pueden no afectar sobre las estimaciones. De hecho, la Tabla 2.2 compara las estimaciones de las probabilidades a posteriori del tamaño de la mixtura para estos casos y se puede observar que las estimaciones son similares.

		1	2	3	4	5	6	7	8	9	10
Caso 2: $P(k \mid \mathbf{x})$	Alg. RJHEr	.000	.005	.194	.260	.217	.145	.088	.052	.026	.009
	Alg. BDHEr	.000	.002	.167	.194	.171	.146	.120	.095	.078	.022
Caso 3: $P(L \mid \mathbf{x})$	Alg. RJMGE	.008	.218	.379	.260	.096	.031	.005	.001	.000	.000
	Alg. BDMGE	.039	.288	.335	.204	.088	.031	.009	.002	.000	.000

Tabla 2.2: Probabilidades a posteriori del tamaño de la mixtura para los datos del caso 2 (arriba) y los del caso 3 (abajo) utilizando los dos tipos de algoritmos y considerando en cada ejemplo el verdadero modelo generador de los datos.

En líneas generales, se puede admitir que en los experimentos que se han realizado con ambos tipos de algoritmos son comparables. La mayor ventaja que se ha encontrado en los algoritmos de tipo BDMCMC desarrollados es que son más sencillos de implementar y que, en algunos casos, pueden visitar estados improbables. Por otro lado, los algoritmos de salto reversible utilizados son menos costosos computacionalmente en cuanto a velocidad y a memoria requerida.

2.5.2.1. Sensibilidad a las elecciones a priori.

Se han realizado diferentes experimentos utilizando distintas distribuciones a priori para el número de términos en la mixtura. Se ha observado que hay poca sensibilidad a estas elecciones en el valor de la moda a posteriori de L y k. Como era de esperar, lo que se aprecia es una concentración mayor de la probabilidad alrededor de la moda a posteriori cuando se utilizan distribuciones Poisson truncadas a priori. Por ejemplo, en el caso 2, si se utilizan distribuciones Poisson de media 2 ó 3 la masa de la distribución a posteriori de k se sigue repartiendo entre los valores $k = 3$ hasta $k = 6$, aunque la probabilidad está más concentrada entre los valores 3 y 4, entre los que se encuentra también la moda a posteriori.

Por otro lado, en el caso de las mixturas de distribuciones Erlang, se ha observado que la información a priori sobre la media del parámetro entero, ν_r influye, en algunos casos, sobre el valor de estos parámetros a posteriori lo que se refleja ligeramente en la estimación de la densidad. En concreto, si se aumenta demasiado el valor, $1/\vartheta$, de la media a priori de ν_r, el algoritmo visitará con menos frecuencia los estados en los que el valor del parámetro ν_r sea pequeño. Consecuentemente, los valores altos de ν_r producirán componentes con varianzas pequeñas ya que la varianza de cada componente, r, es igual a λ_r^2/ν_r, y este efecto se acentuará si el valor de la media, λ_r, de esa componente es pequeño. Sin embargo, se ha observado en la práctica que las componentes que sufren este problema tienen pesos, w_r, muy pequeños y, por tanto, esta sensibilidad no va a afectar demasiado lo cual se puede comprobar observando la función de distribución. Se comenta a continuación este efecto en los ejemplos considerados.

En los casos que se han presentado se ha utilizado la media de ν_r a priori, $1/\vartheta$, igual a 100. Si se aumenta considerablemente el valor de $1/\vartheta$, por encima de 1000, aparecen componentes en las estimaciones con pesos y varianzas pequeñas. En el caso de los datos reales, como se ha mostrado en la Figura 2.8, esta sensibilidad aparece con $1/\vartheta$ igual a 100 donde aparecen varias componentes con varianza muy pequeña y con valores de la media menores de 8. Sin embargo, si se observa la función de distribución, véase la Figura 2.12, el peso de estas componentes es muy pequeño y no va a afectar en las estimaciones de las probabilidades de interés. En la Figura 2.12 se muestra también la función de distribución obtenida al ajustar la distribución MGE que resulta más apropiada para este conjunto de datos.

Para evitar este problema lo natural sería permitir que ϑ fuese un hiperparámetro y construir un modelo jerárquico que evitara estos inconvenientes. Obsérvese que este comportamiento está muy relacionado con la sensibilidad que se aprecia, en otros modelos de mixtura, a la media a priori de la varianza de las componentes, véase Richardson y Green (1997) para mixturas de normales, o Robert et al. (2000) para modelos de cadenas de Markov ocultas. Otros procedimientos para ajustar mixturas de Erlang basados en estimación máximo verosímil, como el propuesto por Asmussen et al. (1996), requieren prefijar de antemano la dimensión de la mixtura y el valor de los parámetros enteros, ν_r. Más comentarios sobre este algoritmo se incluyen en la siguiente Sección, así como una comparación de estimaciones, véase la Figura 2.13.

Con respecto al resto de parámetros de la distribución HEr, $\mathbf{w}$ y $\boldsymbol{\mu}$, no se ha apreciado sensibilidad importante a pequeños cambios en los parámetros de sus distribuciones a priori. Tampoco, se ha apreciado sensibilidad a las elecciones a priori sobre los parámetros de la distribución MGE. Además, para evitar dificultades se estableció un modelo jerárquico sobre las tasas de las fases exponenciales.

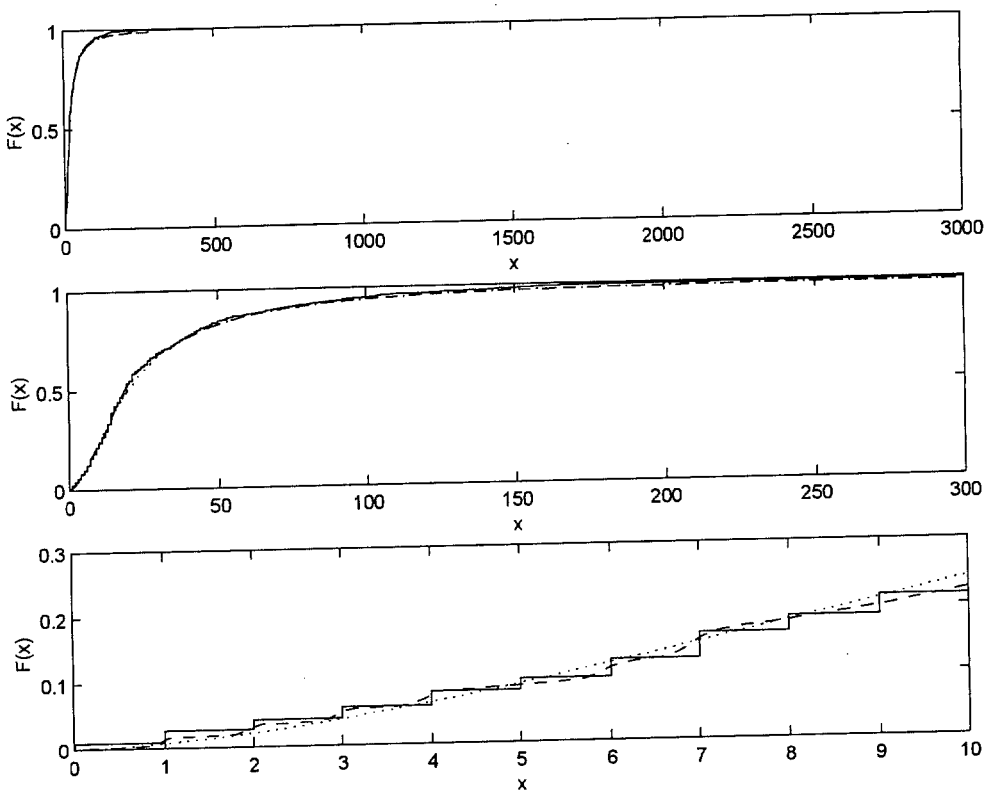


Figura 2.12: Función de distribución empírica para el conjunto de datos reales en el caso 6. Se muestra todo el conjunto de datos reales (arriba), los menores de 300 días (en medio) y los menores de 10 días (abajo). Simultáneamente se muestran las funciones de distribución estimadas asumiendo una distribución HEr (– –) y una distribución MGE (· · ·)

2.6. Comentarios y extensiones.

En este Capítulo, se han desarrollado varios procedimientos para la estimación Bayesiana de densidades continuas y positivas mediante mixturas de distribuciones Erlang y distribuciones Coxianas. Se han construido algoritmos basados en los métodos de salto reversible y en las técnicas BDMCMC permitiendo hacer inferencia sobre los parámetros del modelo de distribución seleccionado, incluida la dimensión paramétrica. También, se ha ilustrado la metodología propuesta con varios ejemplos de datos simulados y un conjunto de datos reales. Estos procedimientos constituyen la herramienta necesaria para ajustar, en los siguientes Capítulos, una muestra de una sucesión de variables aleatorias independientes e idénticamente distribuidas correspondiente a un proceso de servicio o de llegadas a un sistema de colas.

Aunque, en este Capítulo, se han utilizado las distribuciones HEr y MGE como modelos de distribución para un conjunto de observaciones, se podrían suponer otras clases de distribuciones de tipo PH. Incluso, sería posible no asumir ninguna estructura en el modelo y considerar toda la familia de distribuciones de tipo PH, como se hace en Bladt et al. (2003). En esta referencia, se propone un algoritmo MCMC en el que se reconstruye el proceso de Markov implícito en una distribución PH, sin embargo, se supone un número de fases fijo y sería, por tanto, interesante la extensión de este algoritmo a un espacio paramétrico de dimensión desconocida utilizando los métodos RJMCMC y BDMCMC, como se ha procedido en los algoritmos desarrollados en este Capítulo. El inconveniente es que la familia completa de distribuciones de tipo PH es muy general y este hecho puede causar problemas con la identificabilidad de los parámetros que se pretenden estimar. Además, un resultado importante de Cumani (1982) demuestra que un subconjunto considerable de distribuciones de tipo PH, la familia conocida como distribuciones acíclicas, puede representarse de manera única por distribuciones Coxianas, que son las consideradas en este Capítulo, y por tanto, se está considerando ya un subconjunto suficientemente grande de distribuciones de tipo PH que sin embargo incluye un número más reducido de parámetros.

Por otro lado, como se comentó en la Sección 2.5.2, se podrían extender los algoritmos de dimensión variable que se han propuesto utilizando otros tipos de movimientos en los parámetros, por ejemplo, algoritmos en tiempo continuo con movimientos de separación y combinación de componentes. Sería también conveniente modificar el modo de almacenar las muestras en los métodos en tiempo continuo. Por ejemplo, en los algoritmos que se han desarrollado, se observa el proceso en intervalos fijos de tiempo de la misma manera que se propone en Stephens (2000a). Sin embargo, Cappé et al. (2003) proponen un método alternativo para el muestreo del proceso en tiempo continuo que permite reducir varianza de las estimaciones obtenidas a partir de las muestras MCMC y que, básicamente, consiste en guardar los estados que se visitan y el tiempo esperado de permanencia en ellos. Con este procedimiento se puede examinar la proporción de estados que son significativos, en el sentido de que el tiempo de permanencia en ellos es relevante, y comparar estas proporciones con los porcentajes de valores aceptados en los algoritmos de tipo salto reversible ya que, como se observó en la Sección 2.5.2, aunque en los algoritmos en tiempo continuo se aceptan todos los estados que se visitan, no todos ellos influyen significativamente en las estimaciones.

Los métodos expuestos en este Capítulo se pueden comparar con otros procedimientos utilizados para ajustar distribuciones de tipo PH. Por un lado, se han propuesto varias técnicas basadas en el método de los momentos, véase, por ejemplo, Johnson y Taaffe (1991) o Lang y Arthur (1997). Este tipo de aproximación no es efectiva en algunos casos como para las distribuciones con colas pesadas que tienen pocos momentos finitos. Por otro lado, existen varios métodos de estimación máximo verosímil. Uno de los más conocidos es el implementado en Asmussen et al. (1996) donde se utiliza un algoritmo EM para ajustar cualquier distribución de tipo PH. En la Figura 2.13, se muestra la densidad estimada para el conjunto de datos reales introducidos en la Sección 2.5 obtenida a partir del algoritmo EM desarrollado en Asmussen et al. (1996). Para esta estimación se ha utilizado una distribución MGE en la que se ha fijado el número de fases, $L = 5$, previamente. Se compara esta densidad estimada con la obtenida mediante el algoritmo `BDHEr`. Ambas densidades resultan bastante similares ya que se utilizan distribuciones a priori muy poco informativas en el procedimiento Bayesiano. El tiempo de computación para el algoritmo EM, depende del valor de L elegido

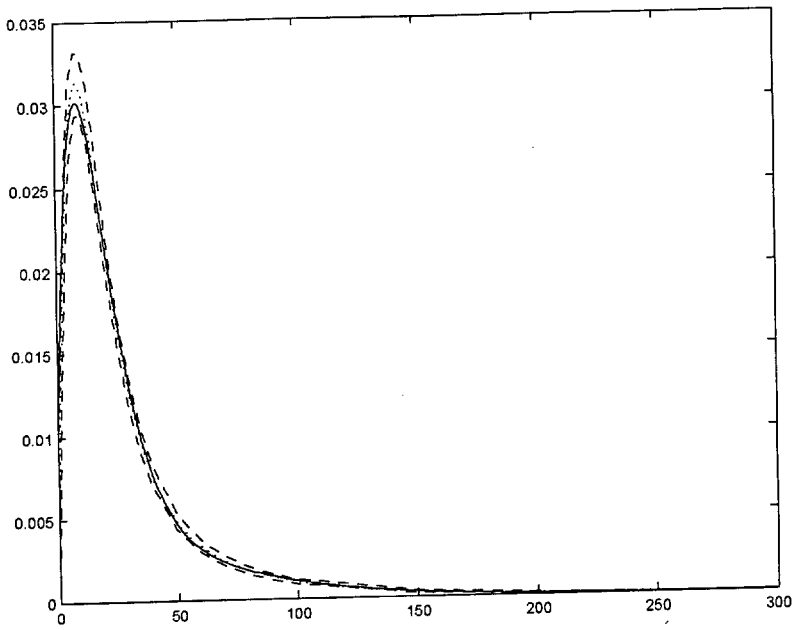


Figura 2.13: Funciones de densidad estimadas con el algoritmo EM (línea continua) y con el algoritmo RJMGE (· · ·) para el conjunto de datos reales del caso 6. Se muestran también el intervalo predictivo al 80 % para a partir del algoritmo RJMGE.

y aunque para $L = 1$, el cálculo es casi inmediato, para $L = 10$, la estimación puede necesitar 20 minutos o más, y por tanto, también son comparables los tiempos globales de computación. Por otro lado, los métodos MCMC de dimensión variable ofrecen más ventajas respecto al algoritmo EM además de no tener que fijar la dimensión de los parámetros a priori. Por ejemplo, al menos bajo nuestro conocimiento, no se han estudiado hasta ahora las propiedades asintóticas de las estimaciones EM de las distribuciones PH. Este hecho dificulta la construcción de intervalos de confianza de los parámetros. Una posibilidad sería el uso de las técnicas bootstrap, sin embargo, actualmente resulta computacionalmente ineficiente puesto que se necesita llevar a cabo muchas estimaciones máximo verosímiles y cada una de ellas requiere un tiempo considerable para su obtención, como se comenta en Asmussen (1997). Como es bien conocido, los métodos MCMC permiten estimar la distribución a posteriori de los parámetros y consecuentemente construir intervalos predictivos de una manera muy sencilla. Asmussen (1997) afirma, además, que los intervalos interesantes en distribuciones PH son los de la densidad y no los de los parámetros, estos intervalos se construyen fácilmente utilizando la muestra MCMC, véase la Figura 2.13. Asimismo, hay que puntualizar que con la metodología Bayesiana, la incertidumbre del modelo se refleja directamente mediante la distribución a posteriori de la dimensión paramétrica, dada en (2.30) y en (2.34), de modo que la densidad predictiva es una media de muchas distribuciones diferentes, véase (2.29) y (2.33).

En esta tesis, se asumen observaciones independientes en el proceso servicio y en el de llegadas. Otra extensión posible sería relajar esta restricción y suponer, por ejemplo, que las llegadas al sistema suceden según un proceso de Poisson modulado de Markov (MMPP) que es, básicamente, un proceso de Poisson cuya tasa varía según un proceso de Markov discreto, es decir, la tasa viene dada por el estado en el que se encuentre el proceso de Markov. Los modelos de colas con este proceso de llegadas han sido ampliamente estudiados, véase, por ejemplo, Neuts (1989). También, existen trabajos en los que se aborda la inferencia para esta situación haciendo uso del algoritmo EM, véase Rydén (1996), y mediante un enfoque Bayesiano, véase Scott y Smyth (2003), sin embargo, en ambos casos se supone que el número de estados del proceso de Markov subyacente es conocido. Se podría generalizar este último procedimiento a casos con dimensión

variable, o también, construir nuevos algoritmos, por ejemplo, en la línea de Bladt et al. (2003) para procesos MMPP. Para ello, se podrían investigar procedimientos similares a los que se desarrollan en Robert et al. (2000) para modelos de cadenas de Markov ocultas (HMM) en los que se hace uso de los métodos de salto reversible.

Finalmente, los procesos de Poisson y los procesos MMPP pueden ser considerados como casos particulares de los procesos de llegadas Markovianos (MAP). El proceso MAP es un modelo útil para caracterizar procesos puntuales que no constituyen procesos de renovación, es decir, se admite correlación entre los tiempos entre-ocurrencias, o entre llegadas si se particulariza el proceso MAP en un proceso de llegadas. La principal propiedad que presenta es que su parametrización matemática lo dota de mucha versatibilidad para describir situaciones que representan procesos de llegadas en muy diferentes contextos. Probabilísticamente está determinado por un generador irreducible infinitesimal A de dimensión $m \times m$, que a su vez es la suma de dos matrices A_0 y A_1. A_1 es una matriz con componentes no negativos y los elementos de la diagonal de A_0 son negativos y no negativos el resto de sus componentes. También es necesario imponer que A_0 tiene inversa y especificar un vector de probabilidades α. Se está usando mucho este proceso en la modelización del tráfico de datos en ordenadores, en Internet y recientemente en la transmisión de datos por vídeo, véase, por ejemplo, Klemm et al. (2003). La diferencia básica con el proceso MMPP es que los cambios de estado del proceso MAP pueden implicar llegadas mientras que en el proceso MMPP sólo se producen llegadas cuando los cambios del proceso modulado son al mismo estado. En Rydén (2000), se recopilan varios métodos de estimación para procesos MMPP y MAP y se argumenta también que la inferencia en procesos MAP tiene una dificultad añadida originada por un problema de identificabilidad en sus parámetros ya que, mientras que para los procesos MMPP se han caracterizado las clases del espacio paramétrico que dan lugar a procesos equivalentes, este problema no se ha resuelto todavía para procesos MAP, lo cual implica que la convergencia en los algoritmos de optimización para la estimación de procesos MAP sea, generalmente, muy lenta. Por tanto, la construcción de métodos asequibles de inferencia para procesos MAP constituye una línea importante de investigación con múltiples aplicaciones en el contexto de sistemas de colas.

Capítulo 3

Análisis Bayesiano del sistema de colas M/G/1 basado en aproximaciones de tipo fase.

En este Capítulo y los siguientes se estudia la estimación de las medidas principales que interesan en un modelo de colas así como la utilización de estas estimaciones en la resolución de algunos de los problemas que se plantean en un sistema concreto, como puede ser, el diseño óptimo y control del mismo. Para estimar estas cantidades de interés, se necesitan los resultados de inferencia sobre el proceso de llegadas y servicio obtenidos en el Capítulo anterior. Por tanto, y para no hacer la lectura tediosa, se supone a partir de ahora que se ha observado el sistema de colas y se tiene información sobre la distribución a posteriori de los parámetros del tiempo de servicio y del tiempo entre las llegadas al sistema. El interés reside ahora en cantidades observables como el número de clientes en la línea de espera, el tiempo de espera en cola y la longitud de los periodos de ocupación.

En concreto, en este Capítulo, se analiza el sistema de colas M/G/1. Tal como se describió en el Capítulo 1, en este modelo de colas los clientes llegan al sistema siguiendo un proceso de Poisson y son atendidos por un único servidor. Los tiempos de servicio son independientes e idénticamente distribuidos según una variable aleatoria general y desconocida. Puesto que se supone que el sistema se ha observado y que, inicialmente, tanto el tiempo entre llegadas como el tiempo de servicio tenían distribuciones desconocidas, en este Capítulo, se supone que, utilizando las técnicas descritas en el Capítulo anterior, las llegadas se ajustan a un proceso Poisson, y que los tiempos de servicio observados se ajustan a uno de los modelos de mixtura propuestos, HEr, o más generalmente, la distribución MGE. Con esta información se pretenden estimar las distribuciones predictivas de las características del sistema de colas M/G/1 observado, utilizando las aproximaciones basadas en los sistemas M/HEr/1 y M/MGE/1. Como se comentó en el Capítulo 1, las distribuciones estacionarias en general existen únicamente si la cola es estable, es decir, si $\rho < 1$, véase (1.6). En este Capítulo se describe cómo examinar si se verifica este requisito una vez observado el sistema.

Para estimar las características predictivas del sistema M/G/1 es necesario conocer previamente la distribución estacionaria asociada a los estados del sistema cuando sus parámetros del sistema son conocidos. Para ello, se puede hacer uso de las herramientas que ofrece la Teoría de Colas clásica donde el modelo de colas M/G/1 ha sido ampliamente estudiado, véase, por ejemplo, Nelson (1995) o Allen (1990). Sin embargo, muchos de los resultados conocidos para este sistema están expresados en términos de transformadas de Laplace, o se conocen, únicamente, los momentos de las distribuciones de interés. Por otro lado, Neuts (1981) introduce una metodología alternativa para el estudio de los modelos de colas basada, fundamentalmente, en la incorporación de técnicas matriciales para caracterizar el sistema de colas. En particular, este enfoque

permite obtener soluciones explícitas de algunas distribuciones estacionarias cuando se aproxima el tiempo de servicio utilizando la clase de distribuciones de tipo PH. Muchos de estos resultados se utilizarán en este Capítulo para analizar los sistemas M/HEr/1 y M/MGE/1, que se incluyen entre los modelos M/PH/1. Las distribuciones de tipo PH surgen como una generalización del método de las etapas (*method of stages*) de manera que muchas de las distribuciones que tienen soluciones explícitas cuando se asumen distribuciones exponenciales se puede obtener cuando se sustituye la distribución exponencial por una de tipo PH, simplificándose, en muchos casos, tediosos cálculos numéricos. Las distribuciones de tipo PH no son aplicables únicamente en para el análisis de sistemas de colas sino que se han utilizado también en otras áreas, como el Análisis de Supervivencia, véase, por ejemplo, Allen (1990), o la Teoría del Riesgo, véase, por ejemplo, Asmussen (2000).

Este Capítulo consta de cinco Secciones. En las dos primeras se asumen conocidos los parámetros del sistema y se incluyen los resultados necesarios de la Teoría de Colas clásica para la inferencia y la predicción que se aborda en la tercera Sección. En concreto, en la Sección 3.1, se incluye una recopilación de los resultados obtenidos por Neuts (1977, 1981) que se utilizarán posteriormente. Esta recopilación contiene, en primer lugar, la definición formal de la clase de distribuciones de tipo PH, así como algunas de sus propiedades y varios ejemplos. A continuación, se incluyen las expresiones matriciales asociadas a las distribuciones estacionarias de algunas características de los sistemas M/PH/1. En la Sección 3.2, se muestra cómo las distribuciones HEr y MGE son de tipo PH y se comprueba que la distribución HEr está incluida en la familia de distribuciones MGE. A continuación, se aplican los resultados de la Sección 3.1 para obtener expresiones explícitas de las características de los sistemas M/HEr/1 y M/MGE/1 cuando los parámetros de la distribución del tiempo entre llegadas y del tiempo de servicio son conocidos. En la Sección 3.3, se utilizan los resultados de las Secciones anteriores para desarrollar inferencia y predicción Bayesiana en un sistema de colas a partir de los datos observados del proceso de llegadas y de servicio. En primer lugar, en la Subsección 3.3.1, se describe cómo examinar si se verifica la condición de ergodicidad en el sistema observado y, a continuación, en la Subsección 3.3.2, se desarrollan procedimientos para estimar las distribuciones predictivas del número de clientes en el sistema, el tiempo de espera en cola y la longitud de los periodos de ocupación en los sistemas que se han propuesto para aproximar el modelo M/G/1, es decir, los sistemas M/HEr/1 y M/MGE/1. Por último, en la Sección 3.4, se ilustra la metodología desarrollada estimando las medidas de interés en sistemas cuyas observaciones del servicio vienen dadas por los conjuntos de datos incluidos en el Capítulo anterior. Por último, la Sección 3.5, se presentan algunos comentarios y extensiones.

3.1. Distribuciones de tipo PH y características del sistema M/PH/1.

En esta Sección, se define la clase de distribuciones de tipo PH introducida por Neuts (1981) y se recopilan algunas de sus propiedades más útiles así como algunos ejemplos de distribuciones PH. También, se recapitulan algunos de los resultados obtenidos por Neuts (1977, 1981) para sistemas de colas M/PH/1 que se caracterizan por tener distribución de servicio de tipo fase y proceso de llegadas Poisson. Estos resultados incluyen expresiones matriciales para las distribuciones estacionarias del número de clientes en el sistema, tiempo de espera en cola y longitud del periodo de ocupación.

El contenido de esta Sección constituye una herramienta fundamental para obtener las cantidades que se presentan en la Sección siguiente y para la inferencia y predicción que se desarrolla en las Secciones posteriores de este Capítulo. Esta Sección presenta únicamente algunos de los resultados que se pueden encontrar en Neuts (1977, 1981) donde se desarrolla un análisis completo sobre esta materia. Para una introducción sobre la familia de distribuciones PH, véase Asmussen (2000, 1987).

3.1.1. Distribuciones de tipo PH.

Si X es una variable aleatoria con soporte en $[0,\infty)$ de tipo PH, entonces, X representa el tiempo hasta la absorción en un proceso de Markov con espacio de estados finito, $E=\{1,...,m+1\}$, donde $m+1$ es el único estado absorbente y los demás estados son transitorios. El generador infinitesimal del proceso viene dado por,

$$Q=\begin{bmatrix} T & \mathbf{T}^0 \\ \mathbf{0} & 0 \end{bmatrix},$$

donde la matriz T de dimensión $m\times m$ verifica que $T_{ii}<0$, para $1\leq i\leq m$, y $T_{ij}\geq 0$, para $i\neq j$. Como la suma de cada fila de Q es cero, entonces $T\mathbf{e}+\mathbf{T}^0=\mathbf{0}$, donde $\mathbf{e}$ es un vector de unos de dimensión $m\times 1$. El vector de probabilidades iniciales de Q es $(\boldsymbol{\alpha},\alpha_{m+1})$ con $\boldsymbol{\alpha}\mathbf{e}+\alpha_{m+1}=1$.

El par $(\boldsymbol{\alpha},T)$ se denomina *representación* de la distribución de X y la dimensión m recibe el nombre de *orden* de la representación. La función de distribución de X viene dada por,

$$F(x)=1-\boldsymbol{\alpha}\exp\{Tx\}\,\mathbf{e},\qquad \text{para } x\geq 0, \tag{3.1}$$

donde $\exp\{Tx\}$ es la exponencial de la matriz Tx que se define por $\exp\{A\}=\sum_{k=0}^{\infty}\frac{A^k}{k!}$ para cualquier matriz cuadrada, A. Consecuentemente, se obtiene que la función de densidad de X es,

$$f(x)=\left\{\begin{array}{ll} \alpha_{m+1} & \text{si } x=0, \\ \boldsymbol{\alpha}\exp(Tx)\mathbf{T}^0 & \text{si } x>0, \end{array}\right\} \tag{3.2}$$

la media de X viene dada por,

$$E[X]=-\boldsymbol{\alpha}T^{-1}\mathbf{e}, \tag{3.3}$$

y su transformada de Laplace-Stieljes, definida en (1.16), viene dada por,

$$f_X^*(s)=\alpha_{m+1}+\boldsymbol{\alpha}(sI-T)^{-1}\mathbf{T}^0,\ \text{para Re}\,s\geq 0, \tag{3.4}$$

donde I es la matriz identidad de orden m y Re denota la parte real.

El conjunto de distribuciones PH definidas en $[0,\infty)$ verifica las dos propiedades siguientes:

- Es una familia cerrada para mixturas finitas. Si $(w_1,...,w_k)$ son los pesos de una mixtura de k distribuciones PH cada una de ellas con representación $(\boldsymbol{\alpha}^r,T^r)$ y orden m^r para cada componente $r=1,...,k$, entonces, la mixtura de las k distribuciones es también PH de orden $\Sigma_{r=1}^{k}m^r$ y con representación $\boldsymbol{\alpha}=\left[w_1\boldsymbol{\alpha}^1,...,w_k\boldsymbol{\alpha}^k\right]$ y $T=diag\left(T^1,...,T^r\right)$, una matriz diagonal por bloques.

- Es una familia cerrada para la convolución de variables. Si X e Y son distribuciones de tipo PH con representaciones $(\boldsymbol{\alpha},T)$ y $(\boldsymbol{\beta},S)$ de órdenes m y n, respectivamente, su convolución es también PH de orden $m+n$ y representación $\boldsymbol{\gamma}=[\boldsymbol{\alpha},\alpha_{m+1}\boldsymbol{\beta}]$ y

$$M=\begin{pmatrix} T & \mathbf{T}^0\boldsymbol{\beta} \\ \mathbf{0} & S \end{pmatrix}.$$

Estas propiedades dotan a la familia de distribuciones PH de mucha flexibilidad, pudiéndose aproximar distribuciones con características muy diferentes. Además, y como propiedad muy importante, se tiene que la clase de densidades PH es densa en el conjunto de densidades definidas en la semirecta real positiva.

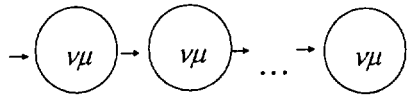


Figura 3.1: Proceso de Markov subyacente para la distribución Erlang.

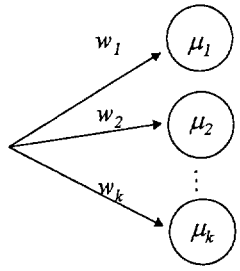


Figura 3.2: Proceso de Markov subyacente en la distribución H_k.

3.1.1.1. Ejemplos de distribuciones PH.

La distribución de tipo PH más sencilla es la distribución exponencial. Una variable distribuida exponencialmente admite una representación de orden $m = 1$ de modo que $(\alpha, T) = (1, -\mu)$, donde $1/\mu$ es la media de la variable.

Otro ejemplo de distribución PH es la distribución Erlang, $Er(\nu, \mu)$, cuya densidad viene dada por (2.2). Obsérvese que la distribución Erlang es la suma de ν fases exponenciales independientes con la misma tasa y por tanto, es la convolución de ν distribuciones PH. La distribución Erlang se puede representar como una distribución PH de orden, ν, con proceso de Markov subyacente que se inicia siempre en el primer estado y visita a todos los demás sucesivamente hasta la absorción que se produce cuando se abandona el último estado, véase la Figura 3.1. Según esta descripción, la representación PH viene dada por el par, (α_{Er}, T_{Er}), donde $\alpha_{Er}=(1,0,...,0)_{(1\times\nu)}$ y

$$T_{Er} = -\nu\mu \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \\ & & & 1 \end{bmatrix}_{(\nu\times\nu)} . \tag{3.5}$$

Otra distribución de tipo PH que se utiliza frecuentemente en la práctica es la mixtura de distribuciones exponenciales o distribución H_k. Su función de densidad viene dada por,

$$f_{H_k}(x) = \sum_{r=1}^{k} w_r \mu_r \exp(-\mu_r x),$$

y admite una representación de orden k con $\alpha_{H_k} = (w_1, ..., w_k)$ y $T_{H_k} = diag(-\mu_1, ..., -\mu_k)$. En este caso, el proceso de Markov implicado puede iniciarse en cualquiera de los estados pero la absorción se produce en el mismo estado en el que comienza el proceso de modo que no se visita ningún estado más, véase la Figura 3.2.

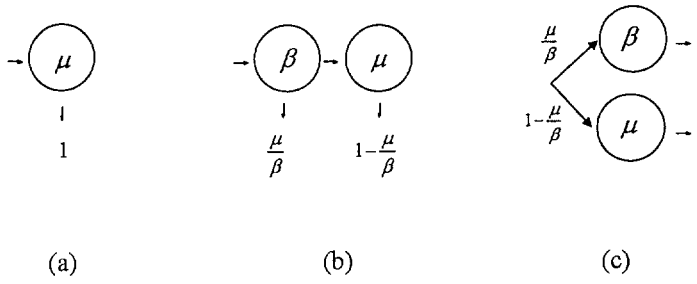


Figura 3.3: Tres procesos diferentes de Markov que representan la distribución exponencial de tasa μ.

3.1.1.2. La ausencia de identificabilidad de las distribuciones de tipo PH.

Una de las características de las distribuciones PH es los parámetros (α, T) no determinan unívocamente a la distribución, es decir, pueden existir varias representaciones de una misma distribución PH, lo cual implica que distintos conjuntos de parámetros pueden dar lugar a la misma distribución. Se presenta a continuación un ejemplo de tres representaciones diferentes para una distribución exponencial de tasa μ que se ilustra en la Figura 3.3. El caso (a) es el más sencillo y corresponde a la representación de orden 1 citada anteriormente. El caso (b) es una representación de orden 2 que da lugar a la distribución exponencial de tasa μ para cualquier $\beta > \mu$. Para comprobarlo, obsérvese que esta representación equivale a la de una distribución MGE dada por,

$$X = \begin{cases} Y_1 & \text{con } P_1 = \frac{\mu}{\beta} \\ Y_1 + Y_2 & \text{con } P_2 = 1 - \frac{\mu}{\beta} \end{cases}$$

donde Y_1 e Y_2 son dos variables exponenciales independientes con tasas β y μ, respectivamente. La función de densidad de X, que se obtiene a partir de (2.4), es la densidad de una exponencial de tasa μ,

$$f(x) = \frac{\mu}{\beta}\beta \exp\{-\beta x\} + \left(1 - \frac{\mu}{\beta}\right)\left[\left(\frac{\mu\beta}{\mu - \beta}\right)\exp\{-\beta x\} + \left(\frac{\mu\beta}{\beta - \mu}\right)\exp\{-\mu x\}\right] = \mu \exp\{-\mu x\}.$$

Por último, es fácil comprobar que la representación del caso (c), que es de orden 2, da lugar a la misma distribución del caso (b) y por tanto, corresponde a una variable exponencial de tasa μ.

El problema de la representación mínima para una distribución de tipo PH, es decir, la dimensión más pequeña posible del espacio de estados de tipo PH, está todavía sin resolver. Neuts (1981) introduce el concepto de representación irreducible que permite, en algunos casos, reducir la dimensión de algunas representaciones. Una representación (α, T) es reducible si la matriz $Q^* = T + (1 - \alpha_{m+1})^{-1}\mathbf{T}^0\alpha$ es reducible. Sin embargo, las representaciones irreducibles tampoco son únicas.

3.1.1.3. Distribuciones de tipo PH discretas.

Se han definido hasta ahora las distribuciones de tipo PH con soporte en $[0, \infty)$. Sin embargo, existe otra familia paralela de distribuciones que es la clase de distribuciones PH discretas. Una variable, X, es de tipo PH discreta de orden m y representación (α, T) si constituye el número de pasos hasta la absorción en una cadena de Markov (en tiempo discreto) con m estados transitorios y uno absorbente y con vector de probabilidades inicial (α, α_{m+1}) y matriz de transición,

$$P = \begin{bmatrix} T & \mathbf{T}^0 \\ 0 & 1 \end{bmatrix},$$

donde $\mathbf{T}^0 = (I - T)\mathbf{e}$, siendo I la matriz identidad de dimensión $m \times m$. La distribución de probabilidad de X viene dada por,

$$P(X = x) = \left\{ \begin{array}{ll} \alpha_{m+1} & \text{si } x = 0 \\ \alpha T^{x-1}\mathbf{T}^0 & \text{si } x = 1, 2, \ldots \end{array} \right\}$$

y la esperanza, $E[X] = \alpha (I - T)^{-1}\mathbf{e}$. El ejemplo más sencillo de distribución PH discreta es la distribución geométrica cuya densidad viene dada por $P(X = x) = p(1-p)^{x-1}$ que admite una representación de orden 1 dada por $(\alpha, T) = (1, 1-p)$. Otros ejemplos de distribuciones PH discretas son la binomial negativa o una distribución degenerada en un valor entero.

3.1.2. Propiedades en equilibrio del sistema de colas M/PH/1.

En esta Subsección, se incluyen varios resultados ya obtenidos en la literatura que permiten calcular las distribuciones estacionarias de algunas características de un sistema de colas M/PH/1. El punto de partida, por tanto, es un sistema con distribución de tiempo entre llegadas exponencial de tasa λ y con distribución de servicio de tipo PH con soporte en $[0, \infty)$ y representación (α, T) de orden m. Se supone además que el sistema está en equilibrio, es decir que la intensidad de tráfico verifica, $\rho < 1$, siendo,

$$\rho = -\lambda \alpha T^{-1}\mathbf{e}.$$

3.1.2.1. Número de clientes en el sistema M/PH/1.

Para obtener la distribución asintótica del número de clientes en el sistema, N, se necesita incorporar la variable aleatoria N_1 definida como el número de llegadas durante un periodo de servicio. En el sistema de colas considerado, N_1, se corresponde con el número de sucesos en un proceso de Poisson de tasa λ durante el intervalo de tiempo que dura la distribución de tipo PH que es independiente del proceso de Poisson. La distribución de N_1 viene dada por,

$$P(N_1 = n) = \int_0^\infty \frac{(\lambda x)^n}{n!} e^{-\lambda x} f_{PH}(x) dx, \tag{3.6}$$

donde $f_{PH}(x)$ representa la densidad del tiempo de servicio. Neuts (1981) demuestra que la distribución de N_1 es de tipo PH discreta con representación (α_{N_1}, T_{N_1}) dada por,

$$\alpha_{N_1} = \lambda \alpha (\lambda I - T)^{-1}, \qquad T_{N_1} = \lambda + (\lambda I - T)^{-1}. \tag{3.7}$$

La distribución de N se puede obtener mediante el siguiente proceso iterativo,

$$\Pr(N = n) = \Pr(N = 0)\Pr(N_1 = n) + \sum_{m=1}^{n+1} \Pr(N = m)\Pr(N_1 = m - n + 1), \qquad \text{para } n \geq 1, \tag{3.8}$$

con $P(N = 0) = 1 - \rho$.

Como se indicó en (1.14), estas ecuaciones son válidas para cualquier sistema M/G/1 con distribución general de servicios, no necesariamente de tipo PH y manteniendo la definición de N_1 dada anteriormente, véase, por ejemplo, Nelson (1995).

3.1.2.2. Tiempo de espera en el sistema M/PH/1.

Se puede demostrar (Neuts (1981)) que la distribución estacionaria del tiempo de espera en cola, W, en un sistema M/PH/1 es de tipo PH con soporte en $[0,\infty)$ y representación $(\boldsymbol{\alpha}_W, T_W)$ dada por,

$$\boldsymbol{\alpha}_W = \rho\boldsymbol{\psi}, \qquad T_W = T + \rho\mathbf{T}^0\boldsymbol{\psi}, \tag{3.9}$$

donde $\boldsymbol{\psi}$ es el vector de probabilidades estacionarias de $T + \mathbf{T}^0\boldsymbol{\alpha}$. Por tanto, $\boldsymbol{\psi}$ es la única solución de las siguientes ecuaciones,

$$\boldsymbol{\psi}(T + \mathbf{T}^0\boldsymbol{\alpha}) = 0, \qquad \boldsymbol{\psi}\mathbf{e} = 1. \tag{3.10}$$

Obsérvese que el orden de la representación PH de W es igual al orden m de la representación de la distribución PH del tiempo de servicio.

3.1.2.3. Periodo de ocupación el sistema M/PH/1.

Neuts (1977) describe cómo obtener la distribución de la longitud del periodo de ocupación en un sistema M/PH/1 en equilibrio. Para ello, es necesario resolver un sistema lineal con un número infinito de ecuaciones diferenciales que se puede truncar con un criterio determinado. Se presenta a continuación la expresión obtenida por Neuts (1977) para la función de distribución del periodo de ocupación. En Secciones posteriores se describirá cómo aproximar su expresión explícita para los sistemas de colas concretos que se consideren.

Sea $B_1(i,j,x)$ la probabilidad de que después de un tiempo x, el periodo de ocupación no haya terminado todavía y que estén $i \geq 1$ clientes presentes en el sistema y la fase del servicio que está operando sea j. Las funciones auxiliares $\mathbf{B}_1(i,x) = [B_1(i,1,x), B_1(i,2,x), ...]$ satisfacen el sistema de ecuaciones diferenciales,

$$\begin{aligned} \frac{d}{dt}\mathbf{B}_1(1,x) &= \mathbf{B}_1(1,x)(T - \lambda I) + \mathbf{B}_1(2,x)\mathbf{T}^0\boldsymbol{\alpha} \\ \frac{d}{dt}\mathbf{B}_1(i,x) &= \mathbf{B}_1(i,x)(T - \lambda I) + \mathbf{B}_1(i+1,x)\mathbf{T}^0\boldsymbol{\alpha} + \mathbf{B}_1(i-1,x) \end{aligned} \tag{3.11}$$

para $i \geq 2$, con $\mathbf{B}_1(1,0) = \boldsymbol{\alpha}, \mathbf{B}_1(i,0) = \mathbf{0}$, para $i \geq 2$.

La función de distribución de la longitud del periodo de ocupación, B, viene dada por,

$$F_B(x) = 1 - \sum_{i=1}^{\infty} \mathbf{B}_1(i,x)\,\mathbf{e}.$$

Se verá más adelante que la distribución de B se puede considerar como la distribución del tiempo hasta la absorción en un proceso de Markov con un número infinito de estados. En concreto, el espacio de estados infinito viene dado por $E = \{(i,j),\ i = 1,2,...,\ 1 \leq j \leq m\}$. El sistema (3.11) se puede truncar considerando que la cola no tiene capacidad infinita sino que existe un valor máximo $n_{\text{máx}}$ para el número de clientes presentes en el sistema de modo que el cliente que llega y encuentra $n_{\text{máx}}$ clientes en el sistema se va sin recibir servicio. El valor máximo se puede fijar según el criterio de Neuts (1977) utilizando la distribución del máximo número de clientes durante un periodo de ocupación. Alternativamente, se ha observado en la práctica que se obtienen buenas aproximaciones si se considera $n_{\text{máx}}$ tal que la probabilidad de que el número de clientes presentes en el sistema supere $n_{\text{máx}}$ sea lo suficientemente pequeña, por ejemplo, $P(N > n_{\text{máx}}) = 0.001$, donde N es la ocupación del sistema dado en (3.8).

3.2. Análisis de los sistemas M/HEr/1 y M/MGE/1.

En esta Sección, se utilizan los resultados recopilados en la Sección anterior para obtener expresiones explícitas sobre las características de interés de los dos sistemas de colas considerados, que son el sistema

M/HEr/1 y, más generalmente, el sistema M/MGE/1, en los que el tiempo de servicio sigue una distribución HEr y MGE, respectivamente.

En primer lugar, se obtienen representaciones PH para las distribuciones HEr y MGE. A partir de estas representaciones se muestra cómo la mixtura de distribuciones Erlang es un caso particular de la distribución MGE, verificando así las afirmaciones del Capítulo anterior. Posteriormente, se obtienen expresiones para las distribuciones estacionarias del tamaño del sistema, el tiempo de espera en cola y la longitud del periodo de ocupación para los dos sistemas de colas mencionados.

A lo largo de toda la Sección, los parámetros del sistema se suponen conocidos. Estos están constituidos por la tasa de llegadas, λ, y los parámetros del servicio, θ_μ, dados por $\theta_\mu = \{k, \mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu}\}$, para el caso en el que se consideren tiempos de servicio distribuidos según el modelo de mixtura HEr, y $\theta_\mu = \{L, \mathbf{P}, \boldsymbol{\mu}\}$, para el caso en el que se considera que el tiempo de servicio sigue una distribución Coxiana. Además, se supone que se verifica la condición de equilibrio, es decir, que la media de la distribución del servicio es menor que λ, y por tanto, existen todas las distribuciones estacionarias que se van a considerar.

3.2.1. Las distribuciones HEr y MGE como distribuciones PH.

Se prueba, en primer lugar, cómo los dos modelos de mixtura considerados HEr y MGE admiten una representación de tipo PH y a continuación, utilizando estas representaciones, se muestra cómo la distribución HEr pertenece a la familia de distribuciones MGE, propiedad que se mencionó sin demostrar en el Capítulo anterior.

La argumentación para las distribuciones HEr proviene de la propiedad especificada en la Sección anterior y consiste en afirmar que las mixturas finitas de distribuciones de tipo PH son también de tipo PH. Entonces, una mixtura de distribuciones Erlang es de tipo PH y admite la siguiente representación, $(\boldsymbol{\alpha}_{HEr}, T_{HEr})$, donde $\boldsymbol{\alpha}_{HEr} = \left[w_1\boldsymbol{\alpha}^1_{Er}, ..., w_k\boldsymbol{\alpha}^k_{Er}\right]$, y

$$T_{HEr} = \begin{bmatrix} T^1_{Er} & 0 & \cdots & 0 \\ 0 & T^2_{Er} & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & T^k_{Er} \end{bmatrix}, \tag{3.12}$$

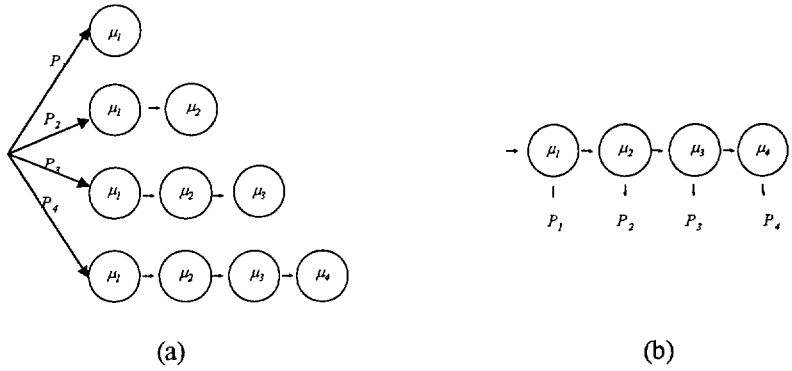
donde $(\boldsymbol{\alpha}^r_{Er}, T^r_{Er})$, es la representación dada en (3.5) de la distribución Erlang de cada componente, $r = 1, ..., k$, de la mixtura. Obsérvese que el orden de $(\boldsymbol{\alpha}_{HEr}, T_{HEr})$ es igual a $\sum_{r=1}^{k} \nu_r$ ya que el orden de la representación de cada distribución Erlang es ν_r.

Además, el valor medio de una distribución HEr viene dado por,

$$\boldsymbol{\alpha}_{HEr} T^{-1}_{HEr} \mathbf{e} = \sum_{r=1}^{k} \frac{w_r}{\mu_r}. \tag{3.13}$$

En el caso de la distribución MGE se considera que es una mixtura de L distribuciones Erlang generalizadas que son a su vez sumas de exponenciales con tasas distintas, véase (2.4). La componente r-ésima de la mixtura, cuya función de densidad viene dada por (2.5), es una distribución de tipo PH ya que es una convolución de r distribuciones exponenciales que son PH y, admite la siguiente representación,

$$\tilde{\boldsymbol{\alpha}}^r_{MGE} = (1, 0, ..., 0)_{(1\times r)}, \qquad \tilde{T}^r_{MGE} = \begin{bmatrix} -\mu_1 & \mu_1 & & \\ & \ddots & \ddots & \\ & & \mu_{r-1} & -\mu_{r-1} \\ & & & \mu_r \end{bmatrix}_{(r\times r)}.$$



(a) (b)

Figura 3.4: Dos representaciones PH diferentes de la distribución MGE. La representación de la izquierda (a) es de orden $0.5L(1+L)$ y corresponde a $(\tilde{\alpha}_{MGE}, \tilde{T}_{MGE})$ mientras que la de la derecha (b) es de orden L y corresponde a (α_{MGE}, T_{MGE})

Por tanto, la distribución MGE es una mixtura de distribuciones de tipo PH, lo cual implica que es también de tipo PH. Según este razonamiento, la distribución MGE admite una representación $(\tilde{\alpha}_{MGE}, \tilde{T}_{MGE})$ donde $\tilde{\alpha}_{MGE} = [P_1\tilde{\alpha}^1_{MGE}, ..., P_L\tilde{\alpha}^L_{MGE}]$, y $\tilde{T}_{MGE} = diag[\tilde{T}^1_{MGE}, ..., \tilde{T}^L_{MGE}]$, una matriz diagonal por bloques. Obsérvese que el orden de la representación $(\tilde{\alpha}_{MGE}, \tilde{T}_{MGE})$ es igual a $\Sigma_{r=1}^{L} r = 0.5L(1+L)$. Sin embargo, se puede probar que la distribución MGE admite una representación de un orden inferior, como se ilustra en la Figura 3.4, y, en este caso, la representación es de orden L y viene dada por,

$$\alpha_{MGE} = (1, 0, ..., 0)_{(1\times L)}, \qquad T_{MGE} = \begin{bmatrix} -\mu_1 & (1-P_1)\mu_1 & & \\ & -\mu_2 & \frac{1-P_1-P_2}{1-P_1}\mu_2 & \\ & & \ddots & \\ & & & -\mu_L \end{bmatrix}_{(L\times L)}. \tag{3.14}$$

Utilizando esta representación, la esperanza de una distribución Coxiana es igual a,

$$\alpha_{MGE} T^{-1}_{MGE} \mathbf{e} = \sum_{r=1}^{L} \frac{1}{\mu_r}\left(1 - \textstyle\sum_{s=1}^{r-1} P_s\right). \tag{3.15}$$

Asmussen (1987) muestra cómo la familia de distribuciones MGE es equivalente a la familia de mixturas de distribuciones Erlang generalizadas, cuya densidad viene dada por (2.5). Con este resultado podemos deducir que la clase de mixturas de distribuciones Erlang está incluida dentro de la distribuciones MGE, ya que las distribuciones HEr están contenidas en las mixturas de distribuciones Erlang generalizadas. La Figura 3.5 ilustra esquemáticamente esta implicación. Para verlo más claro, se muestra a continuación cómo obtener una representación PH de la distribución HEr diferente de la que se presenta en (3.12) que tenga la forma de una representación de una distribución MGE.

Supongamos una distribución HEr con parámetros $\theta = (\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\nu})$. Se asume, sin pérdida de generalidad que $\nu_1\mu_1 = \max\{\nu_1\mu_1, ..., \nu_k\mu_k\}$. Utilizando la representación de una exponencial que se mostró en el caso (b) de la Figura 3.3, se puede representar la primera fase exponencial de tasa $\nu_2\mu_2$ de la segunda componente anteponiendo una fase exponencial de tasa $\nu_1\mu_1$, como se muestra en la Figura 3.6, y análogamente, para el resto de componentes $i = 3, ..., k$. De este modo se obtiene una representación en la que la primera fase es una exponencial de tasa $\nu_1\mu_1$ con probabilidad 1. Procediendo sucesivamente de la misma manera, se consigue una representación como la que aparece en la Figura 3.4 (b) que tiene la estructura dada en (3.14)

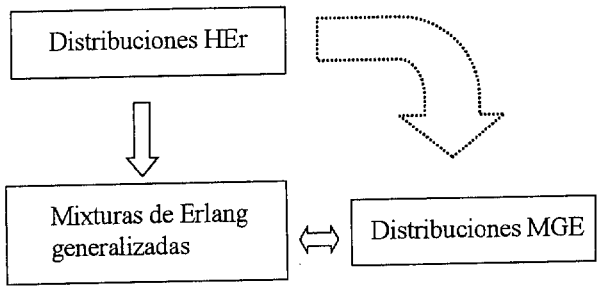


Figura 3.5: Esquema que ilustra el razonamiento por el cual la familia de distribuciones HEr está contenida en la familia de distribuciones MGE.

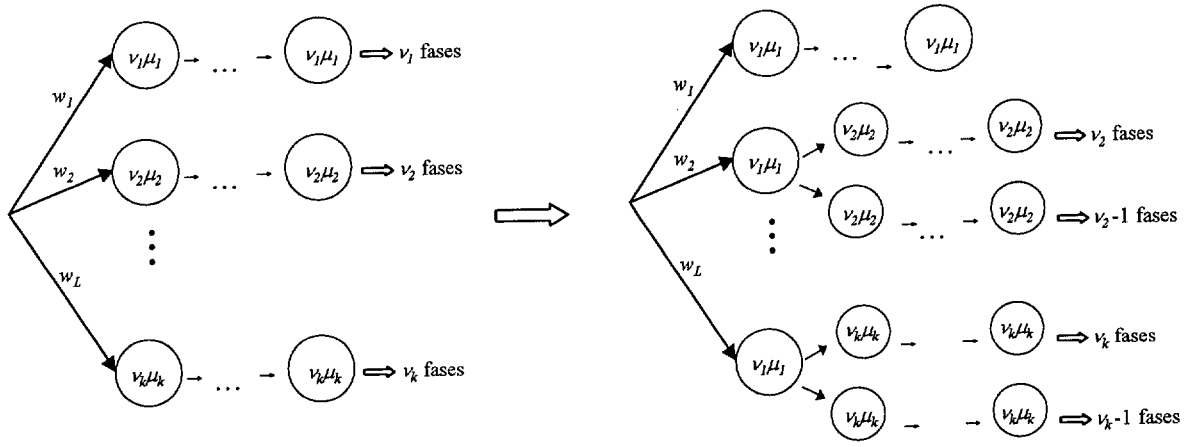


Figura 3.6: Primera etapa para obtener una representación MGE a partir de una representación HEr. Ambos gráficos representan la misma distribución. En el gráfico de la derecha, la primera fase de todas las componentes es una exponencial de tasa $\nu_1\mu_1$.

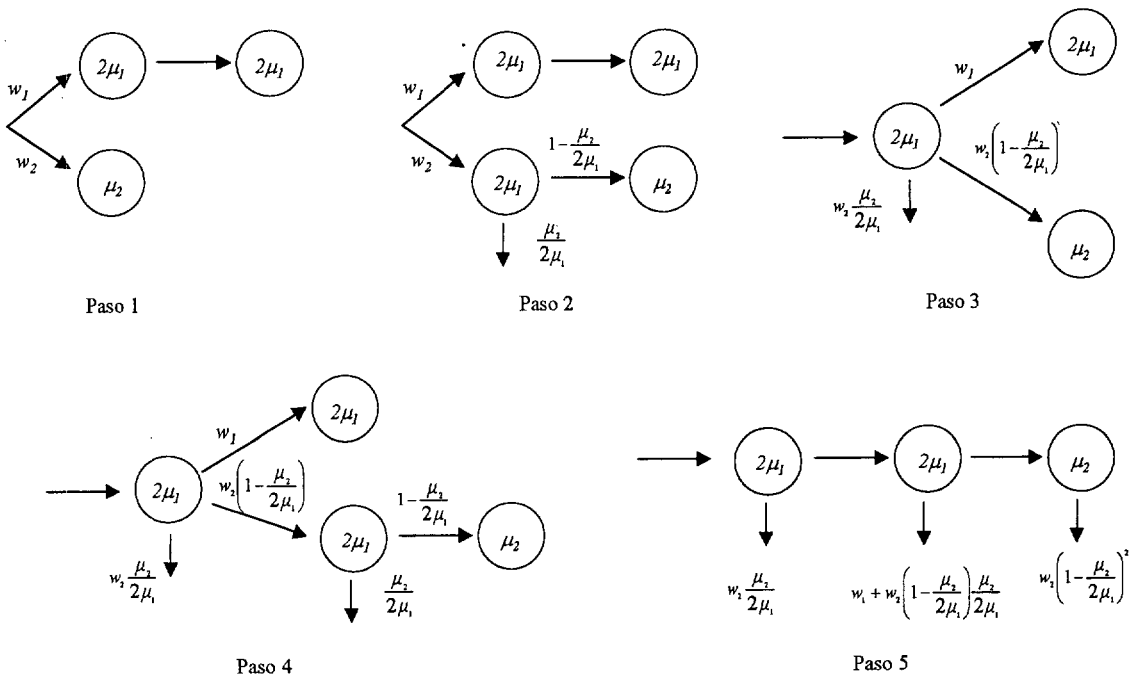


Figura 3.7: Ejemplo de cómo obtener una representación PH de tipo MGE para una distribución HEr con dos componentes. En este ejemplo se asume que $2\mu_1 > \mu_2$ pero se puede obtener una representación análoga si $2\mu_1 > \mu_2$ y una representación de orden 2 si $2\mu_1 = \mu_2$.

para una distribución MGE. La Figura 3.7 muestra un ejemplo concreto de cómo obtener una representación de tipo (α_{MGE}, T_{MGE}) para una mixtura de dos distribuciones Erlang.

Evidentemente, el caso inverso no se verifica, es decir, la familia de distribuciones MGE no es un subconjunto de las distribuciones HEr porque en ese caso las mixturas de distribuciones Erlang generalizadas serían equivalentes a las mixturas de distribuciones Erlang y este hecho no es cierto.

Es importante observar que con este procedimiento el orden de la representación que se obtiene para una distribución HEr es precisamente, $\sum_{r=1}^{k} \nu_r$, que es igual al orden que se obtenía mediante la representación (α_{HEr}, T_{HEr}) dada en (3.12). Por ejemplo, en el la Figura 3.7, el orden de la representación dada originalmente en el Paso 1 es igual a 3 y coincide con el orden que se obtiene en la representación final dada en el paso 5. Por tanto, no se consigue ninguna ventaja en cuanto a la reducción del orden de la representación, es decir, la dimensión de la matriz de la representación es la misma. Por otro lado, como se comentó en el Capítulo anterior, el hecho de desarrollar la inferencia considerando una distribución HEr reduce el tiempo de computación ya que el número de parámetros en menor, lo cual es muy atractivo para alcanzar los objetivos propuestos para estos sistemas de colas.

3.2.2. Número de clientes en el sistema.

La probabilidad de que haya n clientes en un sistema M/HEr/1 ó M/MGE/1 se puede obtener a partir de la ecuación recursiva dada en (3.8) que depende de la distribución del número de clientes, N_1, que llegan al sistema considerado durante un periodo de servicio. La expresión de la distribución de N_1 se podría calcular utilizando la representación PH discreta de N_1 dada en (3.7) para cada sistema. Sin embargo, si se asume que la distribución del servicio es una mixtura de distribuciones Erlang o una distribución Coxiana, se pueden

calcular directamente, como se muestra a continuación.

3.2.2.1. Número de clientes en el sistema M/HEr/1.

En un sistema M/HEr/1, la distribución del número de clientes que llegan al sistema durante un tiempo de servicio es,

$$\begin{aligned} P(N_1 = n) &= \int_0^\infty \frac{(\lambda s)^n}{n!} e^{-\lambda s} f_{HEr}(s \mid \boldsymbol{\theta}_\mu)\, ds \\ &= \sum_{r=1}^{k} w_r \int_0^\infty \frac{(\lambda s)^n}{n!} e^{-\lambda s} \frac{(\nu_r \mu_r)^{\nu_r}}{\Gamma(\nu_r)} s^{\nu_r - 1} \exp(-\nu_r \mu_r s) ds \qquad (3.16) \\ &= \lambda^n \sum_{r=1}^{k} w_r \binom{n+\nu_r-1}{n} \left(\frac{1}{1+\frac{\lambda}{\nu_r \mu_r}}\right)^{\nu_r} (\nu_r \mu_r + \lambda)^{-n}, \qquad (3.17) \end{aligned}$$

donde f_{HEr} es la densidad de una distribución HEr dada en (2.1). En Wiper et al. (2001) aparece un resultado similar para una mixtura de distribuciones gamma, de la cual la distribución HEr es un caso particular.

3.2.2.2. Número de clientes en el sistema M/MGE/1.

Por otra parte, para un sistema M/MGE/1 las probabilidades del número de llegadas durante un tiempo de servicio tienen la expresión siguiente,

$$\begin{aligned} P(N_1 = n) &= \int_0^\infty \frac{(\lambda s)^n}{n!} e^{-\lambda s} f_{MGE}(s \mid \boldsymbol{\theta}_\mu)\, ds \\ &= \sum_{r=1}^{L} P_r \sum_{t=1}^{r} \left(\prod_{s \neq t} \left(\frac{\mu_s - \mu_r}{\mu_s \mu_r}\right)^{-1} \right) \mu_t^{2-r} \int_0^\infty \frac{(\lambda s)^n}{n!} e^{-\lambda s} e^{-\mu_t s} ds \qquad (3.18) \\ &= \sum_{r=1}^{L} P_r \sum_{t=1}^{r} \left(\prod_{s \neq t} \left(\frac{\mu_s - \mu_r}{\mu_s \mu_r}\right)^{-1} \right) \frac{\lambda^n \mu_t^{2-r}}{(\lambda + \mu_t)^{n+1}}, \qquad (3.19) \end{aligned}$$

donde f_{MGE} es la función de densidad de una distribución MGE, dada en (2.4).

3.2.3. Tiempo de espera en cola.

Por los resultados de Neuts (1981) mostrados en la Subsección 3.1.2, la distribución estacionaria del tiempo de espera en cola en los sistemas M/HEr/1 y M/MGE/1 son de tipo PH puesto que, como se ha observado anteriormente, las distribuciones HEr y MGE son de tipo PH.

3.2.3.1. Tiempo de espera en la cola M/HEr/1.

En concreto, si la distribución del servicio es HEr, el tiempo de espera en cola es de tipo PH y admite la siguiente representación de orden $\sum_{r=1}^{k} \nu_r$,

$$\boldsymbol{\alpha}_W = \rho \boldsymbol{\psi}, \qquad T_W = T_{HEr} + \rho \mathbf{T}^0_{HEr} \boldsymbol{\psi}, \qquad (3.20)$$

donde $\boldsymbol{\psi} = (\boldsymbol{\psi}_{\nu_1}, ..., \boldsymbol{\psi}_{\nu_k})$ es el vector de probabilidades estacionario de $T_{HEr} + \mathbf{T}^0_{HEr} \boldsymbol{\alpha}_{HEr}$ y, por tanto, $\boldsymbol{\psi}$ es la solución única a las ecuaciones dadas en (3.10). La expresión explícita de $\boldsymbol{\psi}$ en el sistema M/HEr/1 se

puede obtener a partir de,

$$\psi = -\psi \mathbf{T}^0_{HEr} \alpha_{HEr} T^{-1}_{HEr} = -E[HEr]^{-1} \alpha_{HEr} T^{-1}_{HEr},$$

y se puede comprobar que,

$$\psi_{\nu_r} = \left(\sum_{r=1}^{k} \frac{w_r}{\mu_r} \right)^{-1} \frac{w_r}{\nu_r \mu_r} \mathbf{e}'_{\nu_r},$$

donde $\mathbf{e}'_{\nu_r}$ es un vector de unos de dimensión $1 \times \nu_r$. Por tanto,

$$\alpha_W = \rho\psi = \lambda[\frac{w_1}{\nu_1\mu_1}\mathbf{e}'_{\nu_1}, ..., \frac{w_k}{\nu_k\mu_k}\mathbf{e}'_{\nu_k}],$$

y,

$$T_W = T_{HEr} + \rho \mathbf{T}^0_{HEr} \psi = T_{HEr} + \lambda M,$$

donde M es una matriz por bloques en la que cada bloque M_{ν_r,ν_s} es una matriz $\nu_r \times \nu_s$ tal que,

$$M_{\nu_r,\nu_s} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ w_s \frac{\nu_r \mu_r}{\nu_s \mu_s} & \cdots & w_s \frac{\nu_r \mu_r}{\nu_s \mu_s} \end{bmatrix}.$$

3.2.3.2. Tiempo de espera en la cola M/MGE/1.

Se obtiene a continuación la representación PH de orden L para la distribución del tiempo de espera en cola en un sistema M/MGE/1,

$$\alpha_W = \rho\psi, \qquad T_W = T_{MGE} + \rho \mathbf{T}^0_{MGE} \psi, \tag{3.21}$$

donde ahora ψ resulta ser igual a,

$$\psi = -\psi \mathbf{T}^0_{MGE} \alpha_{MGE} T^{-1}_{HEr} = -E[MGE]^{-1} \alpha_{MGE} T^{-1}_{MGE},$$

de donde se obtiene que,

$$\psi_r = \frac{\left(1 - \sum_{t=1}^{r-1} P_t\right) \frac{1}{\mu_r}}{\sum_{r=1}^{L} \left(1 - \sum_{t=1}^{r-1} P_t\right) \frac{1}{\mu_j}}.$$

Por tanto, la representación de la distribución del tiempo de espera viene dada por,

$$\alpha_W = \rho\psi = \lambda[\frac{1}{\mu_1}, \frac{1-P_1}{\mu_2}, ..., \frac{1-P_1-...-P_L}{\mu_L}],$$

y,

$$T_W = T_{MGE} + \rho \mathbf{T}^0_{MGE} \psi = T_{MGE} + \lambda M,$$

donde M es una matriz $L \times L$ tal que,

$$M_{r,s} = \frac{P_r \mu_r \left(1 - \sum_{t=1}^{s-1} P_t\right)}{\mu_s \left(1 - \sum_{t=1}^{r-1} P_t\right)}.$$

Con estos resultados, se puede obtener una expresión para la función de densidad para la distribución del tiempo de espera en cola, $f_W(x)$, dados los parámetros de un sistema M/HEr/1 ó M/MGE/1,

$$f_W(x) = \begin{cases} 1-\rho, & \text{si } x = 0 \\ \boldsymbol{\alpha}_W \exp\{T_W x\}\, \mathbf{T}_W^0, & \text{si } x > 0 \end{cases} \tag{3.22}$$

donde $\exp\{T_W x\}$ es la exponencial de la matriz $T_W x$ y la función de distribución viene dada por,

$$F_W(x) = 1 - \boldsymbol{\alpha}_W \exp\{T_W x\}\, \mathbf{e}. \tag{3.23}$$

3.2.4. Longitud del periodo de ocupación.

El objetivo en este apartado es obtener una expresión explícita para la distribución de la longitud del periodo de ocupación en los sistemas M/HEr/1 y M/MGE/1. Para ello es necesario resolver el sistema infinito de ecuaciones diferenciales dado en (3.11) que, con el truncamiento considerado en la Subsección 3.1.2, resulta ser igual a,

$$[\mathbf{B}_1'(1,x),...,\mathbf{B}_1'(n_{\text{máx}},x)] = [\mathbf{B}_1(1,x),...,\mathbf{B}_1(n_{\text{máx}},x)] \begin{bmatrix} T-\lambda I & \lambda I & & \\ \mathbf{T}^0\boldsymbol{\alpha} & T-\lambda I & \ddots & \\ & \ddots & \ddots & \lambda I \\ & & \mathbf{T}^0\boldsymbol{\alpha} & T-\lambda I \end{bmatrix}$$

con $[\mathbf{B}_1(1,0),...,\mathbf{B}_1(n_{\text{máx}},0)] = [\boldsymbol{\alpha},\mathbf{0},...,\mathbf{0}]$. Obsérvese que este sistema de ecuaciones diferenciales tiene la estructura siguiente,

$$\boldsymbol{\chi}_B'(x) = \boldsymbol{\chi}_B(x) T_B, \qquad \text{con } \boldsymbol{\chi}_B(0) = \boldsymbol{\alpha}_B, \tag{3.24}$$

donde $\boldsymbol{\chi}_B(x) = [\mathbf{B}_1(1,x),...,\mathbf{B}_1(n_{\text{máx}},x)]$, $\boldsymbol{\alpha}_B = [\boldsymbol{\alpha},\mathbf{0},...,\mathbf{0}]$ y

$$T_B = \begin{bmatrix} T-\lambda I & \lambda I & & \\ \mathbf{T}^0\boldsymbol{\alpha} & T-\lambda I & \ddots & \\ & \ddots & \ddots & \lambda I \\ & & \lambda\mathbf{T}^0\boldsymbol{\alpha} & T-\lambda I \end{bmatrix}. \tag{3.25}$$

La solución del sistema (3.24) es, $\boldsymbol{\chi}_B(x) = \boldsymbol{\alpha}_B \exp\{T_B x\}$. Por tanto, se puede considerar que la distribución de la longitud del periodo de ocupación, B, se aproxima a una distribución de tipo PH con representación $(\boldsymbol{\alpha}_B, T_B)$ cuya función de distribución viene dada por,

$$F_B(x) \simeq 1 - \sum_{i=1}^{n_{\text{máx}}} \mathbf{B}_1(i,x)\,\mathbf{e} = 1 - \boldsymbol{\alpha}_B \exp\{T_B x\}\,\mathbf{e}. \tag{3.26}$$

Este argumento permite considerar la distribución de B en un sistema M/PH/1 como una distribución de tipo PH de orden infinito, es decir, la distribución del tiempo hasta la absorción por el estado absorbente único en un proceso de Markov con infinitos estados transitorios.

3.2.4.1. Periodo de ocupación en un sistema M/HEr/1.

Para el sistema de colas M/HEr/1, la longitud del periodo de ocupación se puede aproximar por una distribución de tipo PH con orden $n_{\text{máx}} \times \sum_{r=1}^{k} \nu_r$ y representación $(\boldsymbol{\alpha}_B, T_B)$ que se obtiene reemplazando

en (3.25) los valores de (α, T) por la representación de del tiempo de servicio (α_{HEr}, T_{HEr}) dada en (3.12) de modo que,

$$\alpha_B = (\alpha_{HEr}, \mathbf{0}, \mathbf{0}, ...), \qquad T_B = \begin{bmatrix} T_{HEr} - \lambda I & \lambda I & & \\ \mathbf{T}^0_{HEr}\alpha_{HEr} & T_{HEr} - \lambda I & \ddots & \\ & \ddots & \ddots & \lambda I \\ & & \mathbf{T}^0_{HEr}\alpha_{HEr} & T_{HEr} - \lambda I \end{bmatrix},$$

donde $\mathbf{0}$ representa un vector de ceros de dimensión $1 \times \sum_{r=1}^{k} \nu_r$ y la matriz I es la matriz identidad de dimensión $\sum_{r=1}^{k} \nu_r \times \sum_{r=1}^{k} \nu_r$. Obsérvese que la matriz $\mathbf{T}^0_{HEr}\alpha_{HEr}$ es una matriz con muchos valores nulos. En concreto, es una matriz por bloques en la que cada bloque $\left(\mathbf{T}^0_{HEr}\alpha_{HEr}\right)_{\nu_r,\nu_s}$ es una matriz $\nu_r \times \nu_s$ tal que,

$$\left(\mathbf{T}^0_{HEr}\alpha_{HEr}\right)_{\nu_r,\nu_s} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ w_s \nu_r \mu_r & 0 & \cdots & 0 \end{bmatrix}.$$

3.2.4.2. Periodo de ocupación en un sistema M/MGE/1.

Análogamente, la representación de la distribución PH que aproxima la distribución de B en un sistema M/MGE/1 es de orden $n_{\text{máx}} \times L$ y se obtiene sustituyendo en (3.25) la representación (α, T) por los valores de (α_{MGE}, T_{MGE}) dados en (3.14). En particular el valor de,

$$\left(\mathbf{T}^0_{MGE}\alpha_{MGE}\right)_{r,1} = \frac{P_r \mu_r}{1 - \sum_{s=1}^{r-1} P_s}, \qquad \text{para } r = 1, ..., L,$$

y $\left(\mathbf{T}^0_{MGE}\alpha_{MGE}\right)_{r,s} = 0$, para $s = 2, ..., L$.

3.3. Inferencia y predicción Bayesiana.

En las Secciones anteriores de este Capítulo se han presentado resultados que describen el comportamiento de sistemas de colas conocidos los parámetros. A partir de esta Sección, los objetivos son diferentes y consisten en la estimación Bayesiana de las medidas interesantes asociadas a los sistemas de colas a partir de los datos observados. Como todos los resultados anteriores se derivan bajo la hipótesis de estabilidad del sistema, previamente, en esta Sección, se describe cómo examinar si se verifica la condición de equilibrio, $\rho < 1$. Este análisis es importante ya que, como se ha comentado anteriormente, si la intensidad de tráfico ρ es menor que 1, entonces el modelo de colas es estable y por tanto, existen las distribuciones estacionarias que se pretenden estimar, en caso contrario, lo que se produce es una saturación del sistema y la convergencia a cero de todas las probabilidades asociadas al sistema, así como una convergencia a infinito de la ocupación del mismo.

Se describe a continuación la situación a partir de la cual se va a desarrollar la inferencia. Se supone que se ha observado un sistema de colas con un único servidor tal y como se describió en el experimento de la Sección 2.1 del Capítulo anterior. Se supone que ajustando uno de los dos modelos de distribución propuestos MGE o HEr a las observaciones de los tiempos entre llegadas, $\mathbf{t} = \{t_1, ..., t_{n_a}\}$, con alguno de los procedimientos descritos en el Capítulo 2, existe evidencia estadística para asumir que el proceso de llegadas es de Poisson. Esta evidencia se obtiene a partir de una elevada probabilidad a posteriori (dada en (2.31) y (2.32)) de que la distribución del tiempo entre llegadas sea exponencial. Se supone también que se ha ajustado uno de los modelos de mixtura considerados a las observaciones de los tiempos de servicio,

$\mathbf{s} = \{s_1, ..., s_{n_s}\}$. Por tanto, se tiene como punto de partida la distribución a posteriori de la tasa de llegadas, λ, dada por, véase (1.20),

$$\lambda \mid \mathbf{t} \sim G\left(a + n_a, b + \sum_{i=1}^{n_a} t_i\right), \tag{3.27}$$

y una muestra de la distribución a posteriori de los parámetros, $\boldsymbol{\theta}_\mu$, de la distribución del tiempo de servicio obtenida a partir de uno de los algoritmos MCMC propuestos. Con esta información, se pueden obtener estimadores consistentes, véase (2.17), de muchas características del sistema. Por ejemplo, el valor de la media del tiempo de servicio, S, se puede aproximar con,

$$E[S \mid \mathbf{s}] \approx \frac{1}{J} \sum_{j=1}^{J} E\left[S \mid \boldsymbol{\theta}_\mu^{(j)}\right]$$

donde,

$$E\left[S \mid \boldsymbol{\theta}_\mu^{(j)}\right] = \sum_{r=1}^{k^{(j)}} \frac{w_r^{(j)}}{\mu_r^{(j)}}, \qquad \text{para } j = 1, ..., J, \tag{3.28}$$

en el caso en el que se haya considerado una distribución HEr para el tiempo de servicio, véase (3.13), y se tenga, por tanto, una muestra MCMC dada por $\{\boldsymbol{\theta}_\mu^{(j)} = \left(k^{(j)}, \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \boldsymbol{\nu}^{(j)}\right)\}_{j=1}^{J}$, o bien,

$$E\left[S \mid \boldsymbol{\theta}_\mu^{(j)}\right] = \sum_{r=1}^{L^{(j)}} \frac{1 - \sum_{s=1}^{r-1} P_s^{(j)}}{\mu_r^{(j)}}, \qquad \text{para } j = 1, ..., J, \tag{3.29}$$

en el caso en el que se haya ajustado el modelo de distribución MGE a los datos del tiempo de servicio, véase (3.15), y se tenga entonces una muestra de la distribución a posteriori de sus parámetros dada por $\{\boldsymbol{\theta}_\mu^{(j)} = \left(L^{(j)}, \mathbf{P}^{(j)}, \boldsymbol{\mu}^{(j)}\right)\}_{j=1}^{J}$.

3.3.1. Inferencia sobre la intensidad de tráfico.

Para examinar si el sistema de colas considerado es estable, se debe analizar la probabilidad a posteriori de que la intensidad de tráfico, ρ, sea menor que 1. Teniendo en cuenta que el valor de la intensidad de tráfico en el sistema considerado viene dado por

$$\rho = \lambda E[S \mid \boldsymbol{\theta}_\mu],$$

entonces, es posible aproximar esta probabilidad con la media de las probabilidades de que λ sea menor que la inversa de la media del tiempo de servicio dada por,

$$P(\rho < 1 \mid \mathbf{t}, \mathbf{s}) = \int_\Theta P\left(\lambda < E[S \mid \boldsymbol{\theta}_\mu]^{-1} \mid \mathbf{t}\right) f(\boldsymbol{\theta}_\mu \mid \mathbf{s})\, d\boldsymbol{\theta}_\mu \approx \frac{1}{J} \sum_{j=1}^{J} F_{\lambda|\mathbf{t}}\left(E\left[S \mid \boldsymbol{\theta}_\mu^{(j)}\right]^{-1}\right), \tag{3.30}$$

donde $F_{\lambda|\mathbf{t}}$ representa la función de distribución de (3.27) y donde $E[S \mid \boldsymbol{\theta}_\mu^{(j)}]$ viene dado en (3.28) o en (3.29), según el caso.

Alternativamente, se puede estimar la probabilidad de que el sistema sea estable utilizando una muestra de la distribución a posteriori de λ, dada en (3.27), y usar la aproximación,

$$P(\rho < 1 \mid \mathbf{t}, \mathbf{s}) \approx \frac{1}{J} \#\left\{\rho^{(j)} < 1\right\}, \tag{3.31}$$

donde,

$$\rho^{(j)} = \lambda^{(j)} E\left[S \mid \boldsymbol{\theta}_\mu^{(j)}\right], \tag{3.32}$$

y donde, $\lambda^{(j)}$ son los valores de la muestra de tamaño J de (3.27) de manera que se obtiene una colección de pares $\{\lambda^{(j)}, \theta_\mu^{(j)}\}_{j=1}^J$ que dan lugar a una muestra de la distribución a posteriori de la intensidad de tráfico. Además, un estimador consistente de la intensidad de tráfico, ρ, es,

$$E[\rho \mid \mathbf{t}, \mathbf{s}] \approx E[\lambda \mid \mathbf{t}] \frac{1}{J} \sum_{j=1}^{J} E\left[S \mid \theta_\mu^{(j)}\right],$$

donde $E[\lambda \mid \mathbf{t}] = (a + n_a)(b + \sum_{i=1}^{n_a} t_i)^{-1}$. También, se puede estimar $E[\rho \mid \mathbf{t}, \mathbf{s}]$ calculando el valor medio de los valores de $\rho^{(j)}$ dados en (3.32).

Si la probabilidad de que el sistema sea estable, estimada a partir de (3.30) o (3.31), es lo suficientemente grande (generalmente, mayor que 0.8), entonces se puede asumir que el sistema es ergódico y por tanto, existen las distribuciones de la características del sistema en el límite, es decir, sus distribuciones estacionarias. Consecuentemente, se puede proceder a la estimación de las mismas, que es el cometido la Sección siguiente. Una estimación de la intensidad de tráfico asumiendo que existe equilibrio en el sistema viene dada por,

$$E[\rho \mid \rho < 1, \mathbf{t}, \mathbf{s}] \approx \frac{1}{R} \sum_{j:\rho^{(j)}<1} \lambda^{(j)} E\left[S \mid \theta_\mu^{(j)}\right], \tag{3.33}$$

donde,

$$R = \#\left\{\left(\lambda^{(j)}, \theta_\mu^{(j)}\right), \text{ tales que } \rho^{(j)} < 1\right\}, \tag{3.34}$$

es el número de pares de la colección $\{\lambda^{(j)}, \theta_\mu^{(j)}\}_{j=1}^J$ que verifican la condición de ergodicidad.

3.3.2. Predicción en el sistema.

En esta Subsección, se asume que el sistema es estable y por tanto, existen las distribuciones estacionarias del número de clientes en el sistema, del tiempo de espera en cola y de la longitud del periodo de ocupación. Se pretende estimar las distribuciones predictivas de estas cantidades utilizando conjuntamente la información que se posee sobre la distribución a posteriori de los parámetros del sistema y las expresiones obtenidas anteriormente para estas distribuciones cuando los parámetros del sistema son conocidos.

Por ejemplo, a partir del conjunto de datos de tiempos entre llegadas y tiempos de servicio, $\{\mathbf{t}, \mathbf{s}\}$, se pueden estimar las probabilidades predictivas del número de clientes en el sistema utilizando la técnica conocida como Rao-Blackwellización, véase Casella y Robert (1996),

$$P(N = n \mid \mathbf{t}, \mathbf{s}, \rho < 1) \approx \frac{1}{R} \sum_{j:\rho^{(j)}<1} P\left(N = n \mid \lambda^{(j)}, \theta_\mu^{(j)}\right) \tag{3.35}$$

donde, R viene dado por (3.34) y es la dimensión del conjunto de pares de la muestra de la distribución a posteriori de los parámetros del sistema que verifican la condición de equilibrio y donde $P(N = n \mid \lambda^{(j)}, \theta_\mu^{(j)})$ se obtiene a partir de las ecuaciones dadas en (3.8) que, a su vez, dependen de la variable N_1 obtenida a partir de (3.17) ó (3.19), según sea el modelo de distribución que se haya utilizado para ajustar la distribución de servicio.

Análogamente, se puede estimar la función de distribución predictiva del tiempo de espera en cola,

$$F_W(x \mid \mathbf{t}, \mathbf{s}, \rho < 1) \approx \frac{1}{R} \sum_{j:\rho^{(j)}<1} F_W\left(x \mid \lambda^{(j)}, \theta_\mu^{(j)}\right), \tag{3.36}$$

donde $F_W(x \mid \lambda^{(j)}, \theta_\mu^{(j)})$ es la función de distribución de W conocidos los parámetros detallada en (3.23) que depende de la representación (α_W, T_W) correspondiente. No obstante, en algunos casos la dimensión

de la matriz, T_W, que es igual al orden de la distribución del tiempo de servicio, puede ser muy elevada. El problema no es muy grande si se supone, por ejemplo, un modelo MGE para el tiempo de servicio y se estiman los parámetros según se describió en el Capítulo anterior, ya que la dimensión de T_W no será nunca superior a 10 que es el valor máximo que se fijó para el número de fases, L. Sin embargo, si se supone una distribución HEr para el tiempo de servicio, la dimensión de T_W, que es $\sum_{r=1}^{k} \nu_r$, puede incrementarse considerablemente para valores elevados de k y de $\{\nu_r,\ r = 1, ..., k\}$, porque aunque k no puede ser mayor que 10, los parámetros enteros, ν_r, no están acotados superiormente y pueden alcanzar valores muy grandes para ajustar, por ejemplo, la varianza pequeña de alguna componente de la distribución del servicio. En estos casos, la evaluación exacta de $\exp\{T_W x\}$ puede resultar muy costosa pero, es posible reducir el tiempo de computación calculando el valor de $\boldsymbol{\alpha}_W \exp\{T_W x\}$ a partir de la solución del siguiente sistema lineal de ecuaciones diferenciales,

$$\boldsymbol{\chi}'_W(x) = \boldsymbol{\chi}_W(x) T_W, \qquad \text{con } \boldsymbol{\chi}_W(0) = \boldsymbol{\alpha}_W. \tag{3.37}$$

Obsérvese que este sistema es análogo al dado en (3.24) construido para calcular el término $\boldsymbol{\alpha}_B \exp\{T_B x\}$ que se utiliza en la aproximación de la función de distribución del periodo de ocupación. Para resolver ambos sistemas, se puede hacer uso, por ejemplo, del método de Runge-Kutta de orden cuatro, véase, por ejemplo, Abramowitz y Stegun (1964).

La función de distribución predictiva de la longitud del periodo de ocupación, B, se puede aproximar del mismo modo que en (3.35) y en (3.36) mediante,

$$F_B\left(x \mid \mathbf{t}, \mathbf{s}, \rho < 1\right) \approx \frac{1}{R} \sum_{j:\rho^{(j)}<1} F_B\left(x \mid \lambda^{(j)}, \theta_\mu^{(j)}\right), \tag{3.38}$$

donde $F_B(x \mid \lambda^{(j)}, \theta_\mu^{(j)})$ es la expresión de la función de distribución dada en (3.26) que se calcula resolviendo el sistema (3.24) para cada valor de la muestra MCMC de los parámetros, es decir,

$$F_B\left(x \mid \lambda^{(j)}, \theta_\mu^{(j)}\right) = 1 - \boldsymbol{\alpha}_B^{(j)} \exp\left\{T_B^{(j)} x\right\} \mathbf{e} = 1 - \boldsymbol{\chi}_B^{(j)}(x)\, \mathbf{e}, \tag{3.39}$$

donde $(\boldsymbol{\alpha}_B^{(j)}, T_B^{(j)})$ y $\boldsymbol{\chi}_B^{(j)}(x)$ son el valor de la representación de B y la solución del sistema (3.24) para cada valor de los parámetros de la muestra MCMC.

Nótese que la dimensión de la matriz T_B en el sistema de ecuaciones (3.24) es igual al número de truncamiento, $n_{\text{máx}}$ multiplicado por el orden de la distribución PH de servicio, es decir, $n_{\text{máx}}$ veces superior a la dimensión de la matriz análoga, T_W, del sistema (3.37) para el tiempo de espera. Por tanto, en los casos en los que el orden del tiempo de servicio sea muy grande, como ocurre cuando el valor de $\sum_{r=1}^{k} \nu_r$ es muy alto en un servicio HEr, se necesitará un tiempo considerable para resolver el sistema (3.24) en cada iteración MCMC y, consecuentemente, será muy costoso calcular la estimación de la distribución predictiva, (3.38). En la práctica, se ha observado que truncando el valor de los parámetros enteros de modo que máx $\{\nu_i\} \leq 50$ ó 100 es suficiente para obtener buenas aproximaciones en la mayoría de las situaciones.

Como se comentó en el Capítulo 1, la estimación de los momentos del tamaño del sistema, del tiempo de espera y del periodo de ocupación es imposible dada la estructura de la distribución a priori que se está considerando. Armero y Bayarri (1994a) y Armero y Conesa (1998) demuestran que si se utilizan distribuciones gamma a priori independientes para las tasas de servicio y de llegadas en la cola M/M/1 y M/Er/1, no existen los momentos de los tiempos de espera y del tamaño de la cola ya que las integrales que se obtienen no convergen. Wiper (1998) concluye que estos momentos tampoco existen para la cola Er/M/1 si la estructura a priori tiene estas características. De hecho, demuestra que si se utilizan distribuciones a priori independientes en las tasas de servicio y llegadas, λ y μ, de una cola G/M/1 tal que la densidad conjunta en $\lambda = \mu$ sea positiva, se obtiene que los momentos de los tiempos de espera en la cola no existen. Para demostrar esta ausencia de momentos, se basa en el hecho de que la esperanza de los tiempos de espera en la cola D/M/1 es la más pequeña de todas las colas G/M/1 dadas las tasas de llegadas y servicios, como aparece, por ejemplo, en Hajek (1983). Del mismo modo, se demuestra que no existen los momentos de las

otras medidas mencionadas. Se puede deducir análogamente que la misma conclusión es cierta para sistemas de colas M/G/1, razonando que la cola M/D/1 tiene la esperanza de los tiempos de espera más pequeña de todas los sistemas M/G/1 dadas las tasas de servicios y llegadas.

3.4. Ilustración numérica.

En esta Sección, se ilustran los procedimientos descritos en este Capítulo para desarrollar inferencia Bayesiana en un sistema de colas M/G/1 a partir de observaciones del proceso de llegadas y de servicio. Para ello, se consideran los conjuntos de datos simulados en la Sección 2.5 del Capítulo anterior como datos de servicio en distintos sistemas de colas. No se utiliza el conjunto de datos reales ya que estas observaciones, que son las duraciones de las estancias de enfermos en un hospital, están asociadas a los tiempos de servicio de un sistema de colas con más de un servidor de modo que, el número de servidores se corresponde con el número de camas en el hospital. Este conjunto de datos se estudiará en el Capítulo 3.

Se consideran, por tanto, cinco sistemas de colas con un único servidor. Se supone que se observa para cada sistema un conjunto de 100 datos de servicio correspondientes a los datos simulados de las distribuciones exponencial, HEr, MGE, degenerada y Weibull, de la Sección 2.5 del Capítulo anterior. Para cada sistema, se considera el par de muestras MCMC, obtenidas en la Sección 2.5, de la distribución a posteriori de los parámetros del servicio ajustando una distribución HEr y una distribución MGE, resultantes de los algoritmos `RJHEr` y `RJMGE`, respectivamente. Por simplicidad, se asume que la tasa de llegadas, λ, es conocida y es tal que, la intensidad de tráfico, ρ, es la misma para los cinco sistemas e igual a 0.6.

$P(\rho < 1 \mid \text{datos})$	Caso 1: M/M/1	Caso 2: M/HEr/1	Caso 3: M/MGE/1	Caso 4: M/D/1	Caso 5: M/Weib/1
Alg. RJHEr	0.9980	0.9995	0.9881	0.9964	0.9973
Alg. RJMGE	0.9990	0.9994	0.9971	0.9891	0.9962

Tabla 3.1: Probabilidad a posteriori de que cada uno de los sistemas considerados sea estable para cada una de las dos muestras MCMC resultantes de los algoritmos `RJHEr` y `RJMGE`.

La Tabla 3.1 muestra, para los cinco modelos de colas, la estimación de la probabilidad a posteriori de que el sistema sea estable, $P(\rho < 1 \mid \mathbf{t}, \mathbf{s})$, utilizando la aproximación dada en (3.31). Se muestra el resultado de esta estimación utilizando las dos muestras MCMC de la distribución a posteriori de los parámetros de servicio considerando una distribución HEr y una distribución MGE. Se observa que, en todos los casos, esta probabilidad es bastante grande, mayor que 0.95, y por tanto, parece razonable suponer que se verifica la condición de equilibrio y, consecuentemente, se pueden estimar las distribuciones estacionarias de las medidas del sistema consideradas. Nótese, también, que las probabilidades obtenidas para cada sistema son similares utilizando cualquiera de los dos algoritmos `RJHEr` ó `RJMGE`. La Tabla 3.2 muestra la estimación de la media de ρ a posteriori asumiendo equilibrio. En todos los casos esta estimación está próxima al verdadero valor que es 0.6.

$E[\rho \mid \rho < 1, \text{datos}]$	Caso 1: M/M/1	Caso 2: M/HEr/1	Caso 3: M/MGE/1	Caso 4: M/D/1	Caso 5: M/Weib/1
Alg. RJHEr	0.6279	0.6062	0.6010	0.6033	0.5877
Alg. RJMGE	0.6177	0.6093	0.5963	0.5813	0.5762

Tabla 3.2: Esperanza a posteriori de la intensidad de tráfico para cada uno de los sistemas asumiendo equilibrio y para cada una de las dos muestras MCMC resultantes de los algoritmos `RJHEr` y `RJMGE`.

La Figura 3.8 muestra, con dos tipos de líneas discontinuas, las estimaciones de las probabilidades estacionarias del tamaño, N, de cada uno de los sistemas, obtenidas utilizando (3.35) para las muestras MCMC resultantes de los algoritmos RJHEr y RJMGE. Obsérvese que la estimación de la probabilidad de que el sistema esté vacío, que es $(1-\rho)$ en cualquier modelo de colas M/G/1, es igual a uno menos el valor medio a posteriori de ρ que se indica en la Tabla 3.2, es decir,

$$P(N=0 \mid \rho<1, \text{datos}) = 1 - E[\rho \mid \rho<1, \text{datos}].$$

Las estimaciones en líneas discontinuas se comparan con los valores teóricos de la distribución del número de clientes en el sistema, dados los parámetros, que se presentan en línea continua. En los cinco casos, las estimaciones son muy similares a los valores verdaderos. A continuación se describe cómo obtener la distribución verdadera de N en cada uno de los cinco sistemas. Conocidos los parámetros, para los casos 1 a 3, se puede calcular la distribución de N mediante las ecuaciones dadas en (3.8) y las expresiones dadas en (3.17) y (3.19). En particular, la distribución de N en un sistema M/M/1 es geométrica y viene dada por (1.10). La distribución de N en el caso 4, que considera el sistema M/D/1, viene dada por, véase Nelson (1995),

$$\Pr(N=n) = (1-\rho)\sum_{m=0}^{n} e^{m\rho}(-1)^{n-m}\frac{(m\rho+n-m)(m\rho)^{n-m-1}}{(n-m)!}, \tag{3.40}$$

y por último, la distribución de N del sistema M/Weib/1, en el caso 5, se puede obtener a partir de las ecuaciones dadas en (3.8) donde la distribución del número de clientes que llegan al sistema durante un periodo de servicio, N_1, viene dado por,

$$P(N_1=n) = \int_0^\infty \frac{(\lambda x)^n}{n!}e^{-\lambda x} f_{Weib}(x)\,dx = \int_0^\infty \frac{(\lambda x)^n}{n!}e^{-\lambda x} abx^{b-1}\exp\left\{-ax^b\right\}dx,$$

cuyo valor se puede aproximar mediante un método de integración numérica. En la ilustración se ha utilizado el método de la cuadratura de Lobatto, véase Gander y Gautschi (2000), implementado en MATLAB. También, en la Figura 3.8, se puede observar que se obtienen valores muy parecidos si se supone una distribución HEr para el tiempo de servicio o el modelo de distribución más general MGE.

En la Figura 3.9, se ilustran las estimaciones de la función de distribución del tiempo de espera en cola, W, en cada uno de los sistemas calculadas utilizando (3.36). Obsérvese que el valor de estas funciones en el origen corresponde a la probabilidad de que el tiempo de espera sea nulo y coincide con la probabilidad de que el número de clientes presentes en el sistema sea 0, que se muestran en la Figura 3.8. En la Figura 3.9, se indica, como antes, con dos tipos de líneas discontinuas, las funciones obtenidas con cada una de las dos muestras MCMC consideradas. Se ilustran también en línea continua las funciones de distribución de cada uno de los sistemas conocidos los parámetros. Su expresión es fácil de obtener en los casos 1 a 3 ya que la distribución de W es de tipo PH, véase (3.20) y (3.21). En particular, el tiempo de espera en cola en el sistema M/M/1 es proporcional a una distribución exponencial, véase (1.11). En los casos 2 y 3, se observa que la distribución predictiva se aproxima mejor a la función verdadera si se considera el verdadero modelo generador de los datos de los tiempos de servicio, es decir, en el caso 2, la estimación es mejor si se utiliza la muestra del algoritmo RJHEr y, en el caso 3, si se utiliza la muestra del algoritmo RJMGE, aunque las estimaciones son similares con cualquiera de las dos muestras MCMC. En el caso 4, la verdadera función de distribución de W se aproxima bien con cualquiera de las dos distribuciones predictivas a pesar de que, en este caso, la función de distribución verdadera no es diferenciable. Este hecho se puede comprobar teniendo en cuenta que, en el sistema M/D/1, el tiempo de espera en cola de un cliente que llega y encuentra n clientes en el sistema es una uniforme definida en el intervalo $[\frac{n-1}{\mu}, \frac{n}{\mu}]$, donde $1/\mu$ es el tiempo degenerado de servicio. Condicionando a todos los posibles valores del número de clientes, n, se obtiene que la función de distribución de W tiene la siguiente expresión,

$$F_W(x) = 1-\rho+\mu P\left(N \leq \text{máx}\left\{n : x < \frac{n}{\mu}\right\}\right),$$

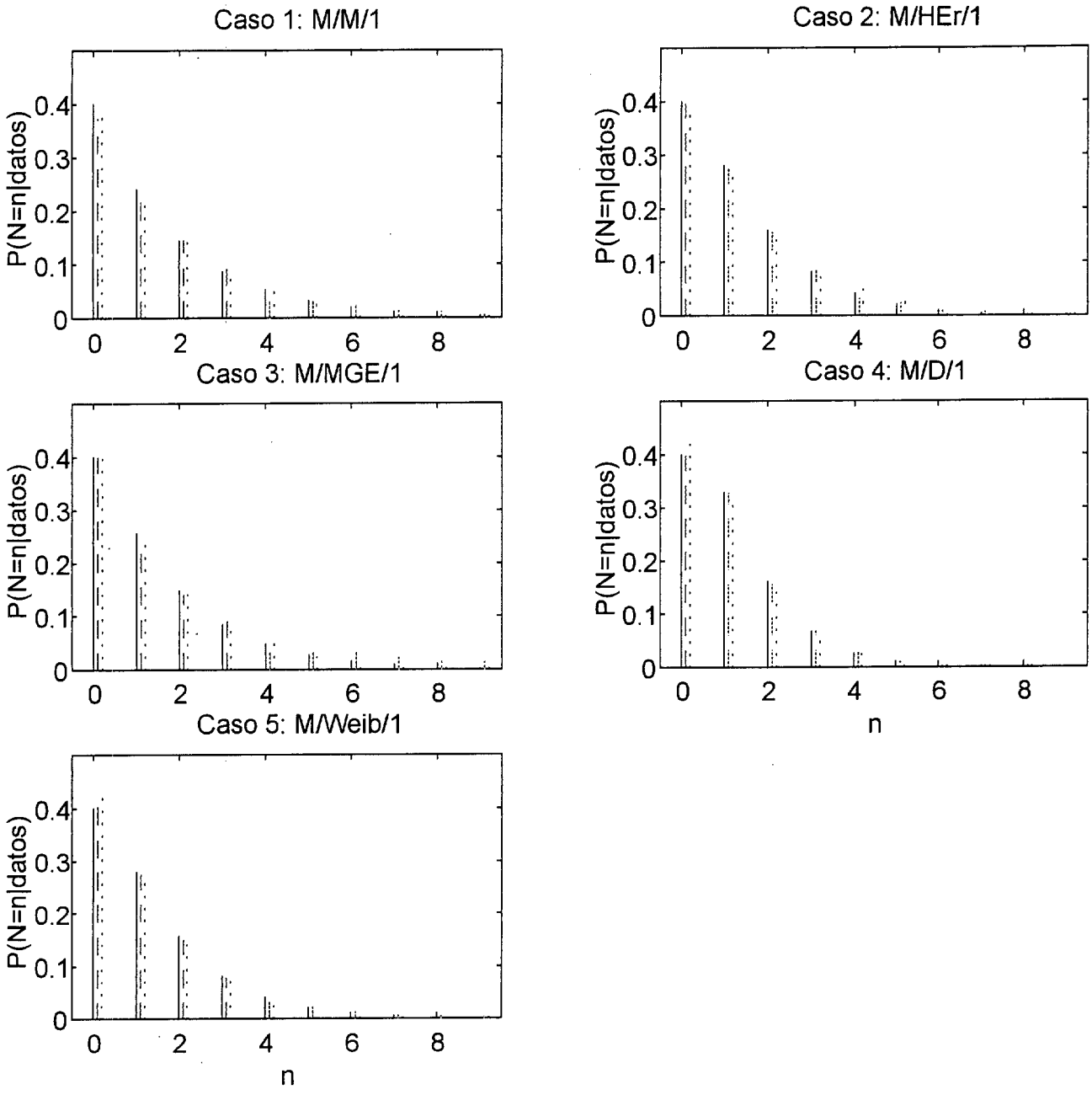


Figura 3.8: Probabilidades predictivas a posteriori del tamaño del sistema utilizando la muestra MCMC del algoritmo `RJHEr` (——) y la del algoritmo `RJMGE` (· · ·) y las probabilidades verdaderas (–).

donde N es el número de clientes en el sistema M/D/1 dado en (3.40). Aunque en la Figura 3.9 no se puede apreciar fácilmente, esta función es únicamente diferenciable a trozos. Para el caso 5, no se presenta, en la Figura 3.9, el valor verdadero de la función de distribución de W, que debería aparecer en línea continua, ya que no se conoce hasta ahora una expresión explícita para esta distribución en el sistema M/Weib/1 conocidos los parámetros.

Las Figuras 3.10 y 3.11 ilustran con los dos tipos de líneas discontinuas varias funciones de distribución de la longitud del periodo de ocupación, B, obtenidas mediante (3.39), para diferentes valores de los parámetros dados en unas observaciones escogidas al azar de cada una de las dos muestras MCMC resultantes de los algoritmos `RJHEr` y `RJMGE`, con diferentes tamaños de la mixtura, k y L, respectivamente. Se ilustra también con línea continua la verdadera función de distribución en todos los casos menos en el 5 que corresponde al sistema M/Weib/1 para el que, nuevamente, se desconoce su expresión explícita. Conocidos los parámetros, el valor verdadero de la función de densidad de B en el caso 1 depende de una función de Bessel modificada y viene dada en (1.13). En los casos 2 y 3, como se describió en la Sección 3.2.4, la verdadera función de distribución se puede aproximar con la representación PH dada en (3.25). En el caso 4, correspondiente al sistema M/D/1, la duración de un periodo de ocupación es siempre un múltiplo de la duración constante del tiempo de servicio y por tanto la distribución de B es discreta. Su expresión se puede obtener teniendo en cuenta que la distribución del número de clientes servidos durante un periodo de ocupación, N_2, viene dada por, véase Kleinrock (1975),

$$\Pr(N_2 = n) = \frac{(n\rho)^{n-1}}{n!} e^{-n\rho},$$

y por tanto, la probabilidad de que el periodo de ocupación sea n veces el tiempo de servicio es igual a la probabilidad de que lleguen n clientes durante un periodo de ocupación. Se puede observar que en todos los casos las estimaciones de la distribución son similares a la verdadera distribución discreta. La diferencia más significativa entre las estimaciones con las muestras `RJHEr` y `RJMGE` se aprecia en el sistema M/D/1. El motivo es que la distribución HEr puede aproximar favorablemente la distribución degenerada de servicio con valores muy grandes del parámetro entero ν lo que conduce a un número elevado de fases, mientras que el modelo de distribución MGE que se ha considerado no permite más de 10 fases, véase la Sección 2.5. A cambio, en este caso, la estimación con la muestra `RJHEr` es mucho más costosa computacionalmente que con la muestra `RJMGE`, como se comenta a continuación.

Todos los cálculos de esta Sección se ha llevado a cabo utilizando MATLAB. Los tiempos de computación para estimar la distribución del tamaño del sistema son similares si se utiliza la muestra MCMC del algoritmo `RJHEr` o si se utiliza la muestra obtenida a partir del algoritmo `RJMGE` y, en ningún caso, son superiores a cinco minutos. Esta similitud en el coste computacional no se aprecia en la estimación la distribución del tiempo de espera en cola, W, y del periodo de ocupación, B, ya que el tiempo requerido con la muestra del algoritmo `RJMGE` es, en todos los casos, mucho menor que el que se necesita con la muestra del algoritmo `RJHEr`. El motivo fundamental es que para aproximar ambas distribuciones estacionarias es necesario resolver sistemas de ecuaciones cuya dimensión depende del orden de la distribución de servicio, que es L en el caso MGE, y por tanto, inferior a 10, y $k \times \Sigma_{r=1}^{k} \nu_r$ en el caso HEr, que no está acotado superiormente, véase la Sección 3.3.2. Concretamente, las estimaciones de las funciones de distribución de W y de B en un sistema determinado requieren entre 10 y 15 minutos con una muestra de `RJMGE`, mientras que pueden ser necesarias entre 2 y 3 horas con una muestra `RJHEr`. En este último caso, los tiempos de computación pueden variar bastante dependiendo del valor de $\max\{\nu_r\}$. Como se comentó en la Sección 3.3.2, en todos los ejemplos se ha truncado en $\max\{\nu_r\} = 100$, a excepción del periodo de ocupación del sistema M/D/1, en la Figura 3.10, donde se consideró $\max\{\nu_r\} = 1000$. El valor de $n_{\max}$ para aproximar B se ha considerado, en todos los casos, tal que la probabilidad de que haya más de $n_{\max}$ clientes presentes en el sistema sea menor que 0.001, véase la Sección 3.1.2.

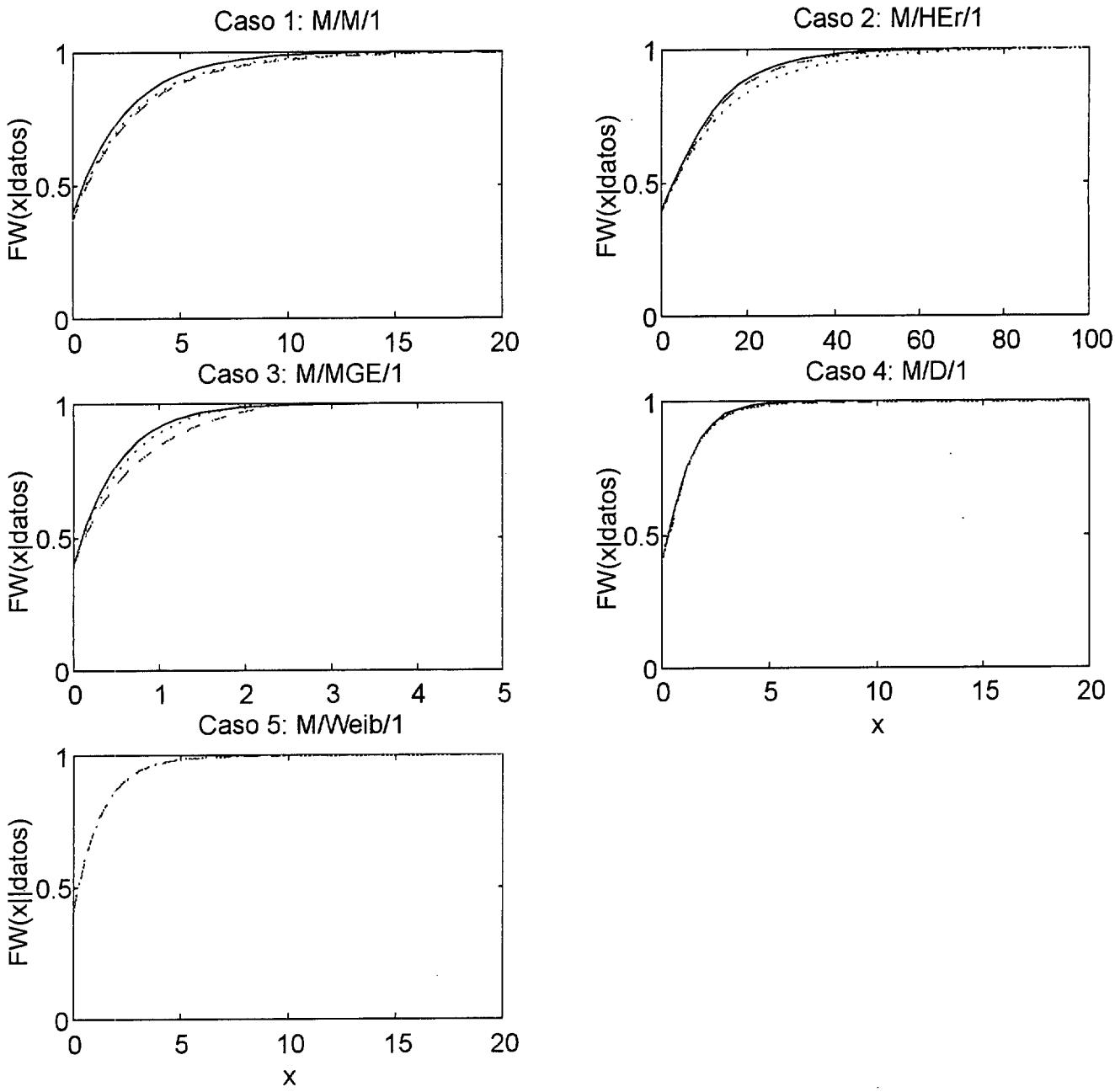


Figura 3.9: Funciones de distribución predictivas del tiempo de espera en cola utilizando la muestra MCMC del algoritmo RJHEr (−−) y la del algoritmo RJMGE (···) y las funciones de distribución verdaderas (−).

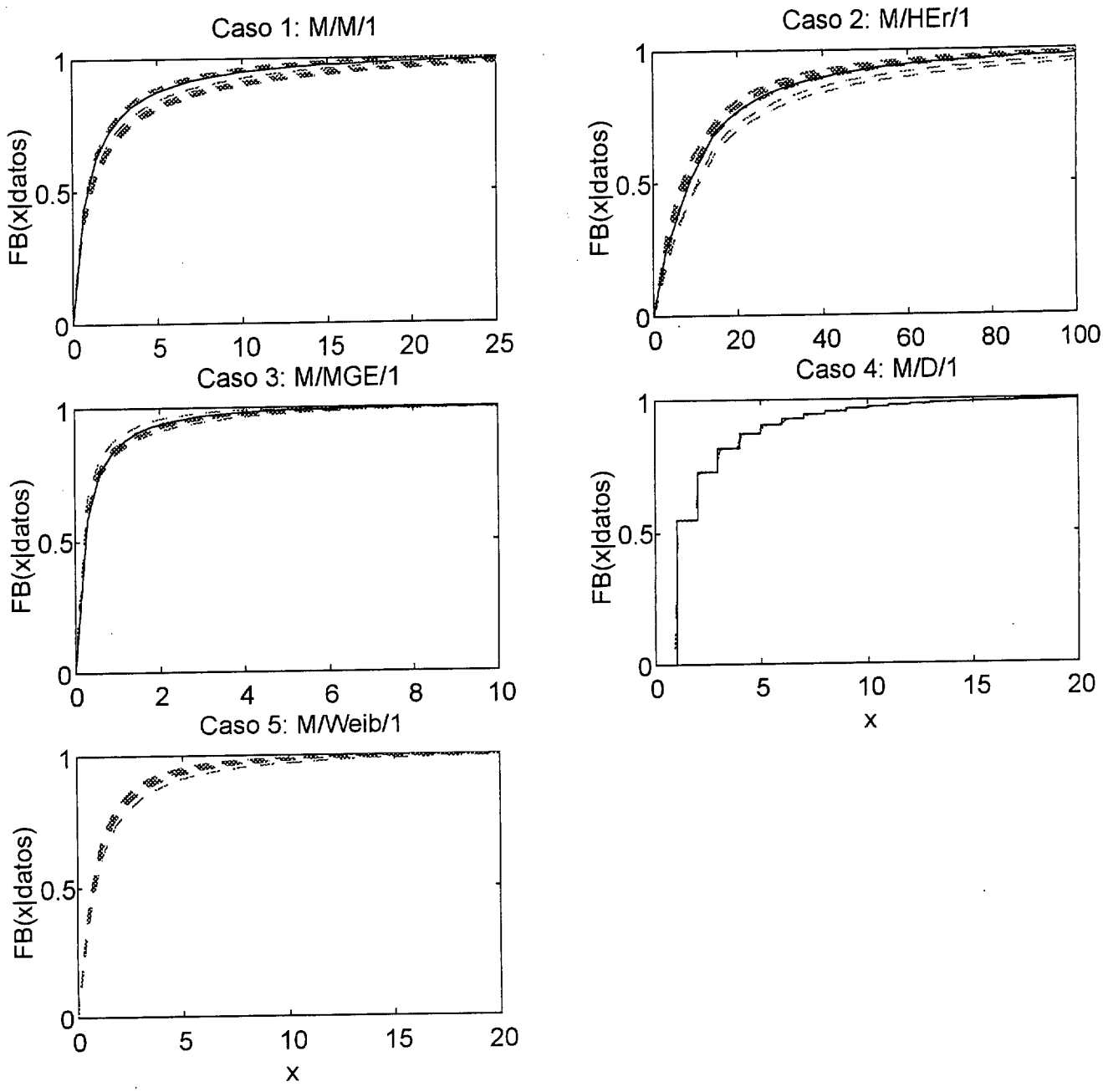


Figura 3.10: Funciones de distribución predictivas de la longitud del periodo de ocupación utilizando la muestra MCMC del algoritmo `RJHEr` (——) y las funciones de distribución verdaderas (−).

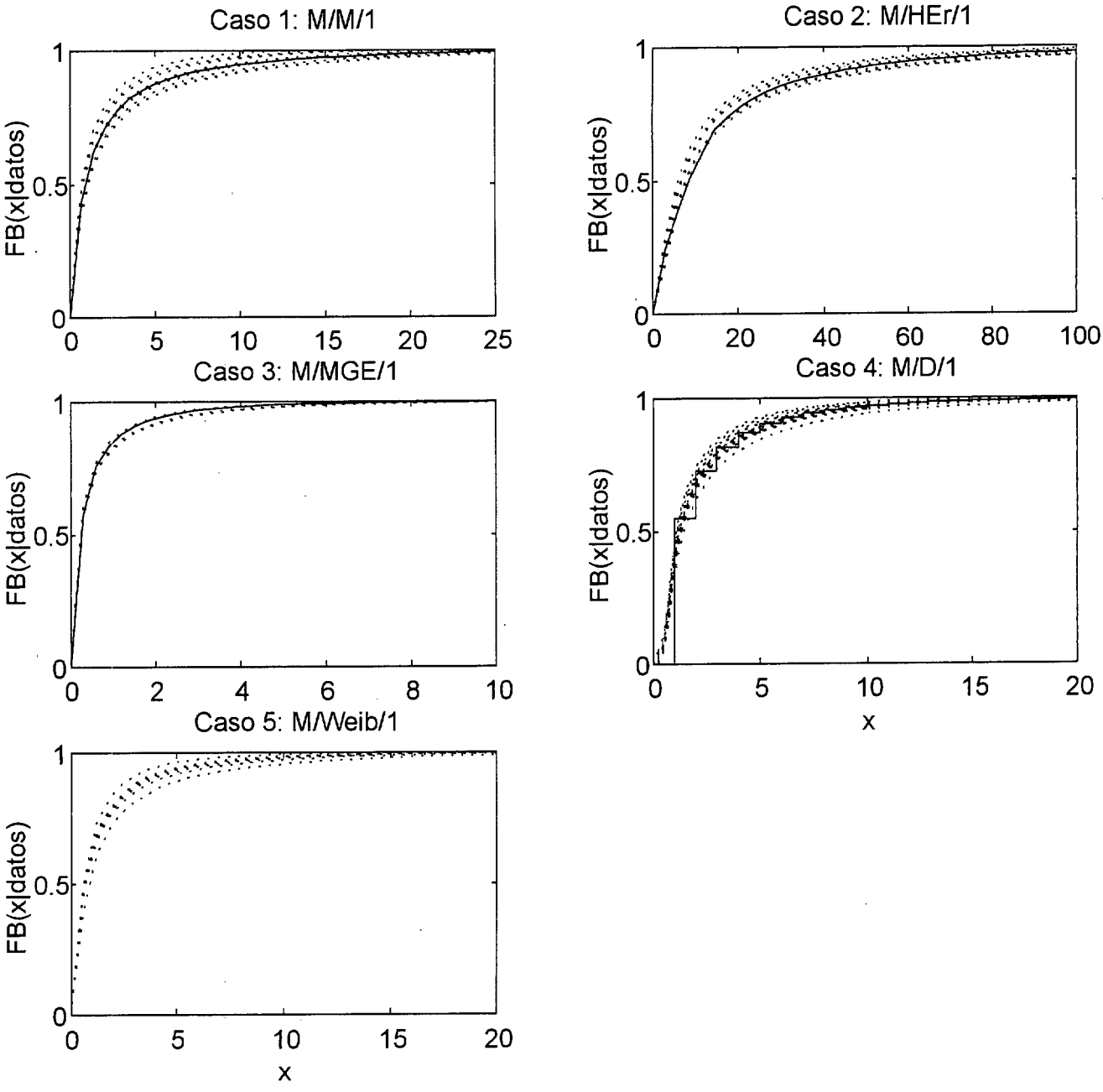


Figura 3.11: Funciones de distribución predictivas de la longitud del periodo de ocupación utilizando la muestra MCMC del algoritmo `RJMGE` $(\cdots)$ y las funciones de distribución verdaderas $(-)$.

3.5. Comentarios y extensiones.

En este Capítulo, se ha desarrollado un análisis Bayesiano de los sistemas de colas M/G/1 considerando las mixturas de distribuciones Erlang y las distribuciones Coxianas para modelar la distribución general del tiempo de servicio. Se ha mostrado que estos modelos de mixtura pertenecen a la clase de distribuciones de tipo PH y que, en particular, la familia de distribuciones HEr está contenida en el conjunto de distribuciones MGE. Se han utilizado las propiedades de las distribuciones de tipo PH para obtener expresiones de las distribuciones estacionarias de algunas características del sistema cuando los parámetros son conocidos. La combinación de estas expresiones con los algoritmos MCMC del Capítulo anterior ha permitido estimar las distribuciones predictivas de estas cantidades de interés que son, el número de clientes en el sistema, el tiempo de espera en cola y la longitud del periodo de ocupación. Finalmente, se ha ilustrado el procedimiento con varios ejemplos.

Aunque teóricamente la distribución HEr es un caso particular de un modelo MGE, el conjunto de modelos HEr considerados para la inferencia no está contenido en el conjunto de modelos MGE. El motivo es que se está suponiendo a priori que el valor máximo para el número de fases L de la distribución MGE es 10 mientras que muchos modelos de distribución HEr considerados requieren más de 10 fases para poder representarse como una distribución MGE. En muchos casos, como se ilustró en el Capítulo anterior, el modelo de distribución HEr considerado permite reducir el coste computacional para estimar favorablemente la distribución del tiempo de servicio. Sin embargo, en este Capítulo, se ha concluido que el hecho de permitir un número elevado de fases en el tiempo de servicio conduce a un incremento considerable en los tiempos de ejecución de los algoritmos que permiten estimar las características del sistema de colas. En estos casos, el modelo de distribución MGE considerado da lugar a estimaciones bastante similares, en la mayoría de los casos, a las obtenidas con la distribución HEr y mucho menos costosas. Además, con un tiempo de servicio MGE, la obtención de las expresiones explícitas de las distribuciones estacionarias es más sencilla, así como su implementación en un algoritmo MCMC.

El hecho de que los modelos de mixtura que se han propuesto, las distribuciones HEr y MGE, sean de tipo PH permite aplicar muchos otros resultados clásicos y recientes de la Teoría de Colas, aparte de los considerados en este Capítulo, para desarrollar inferencia y predicción en sistemas de colas. Por ejemplo, es conocido que la distribución estacionaria del tiempo de espera, W, en un sistema de colas GI/PH/1 es también de tipo PH, véase Asmussen (1992). Además, el número de fases de W es el mismo que el de la distribución del tiempo de servicio. Sin embargo, la representación de W viene caracterizada en términos de una matriz que debe obtenerse iterativamente como la solución de un problema de punto fijo y consecuentemente, si se combinan estos algoritmos con los métodos MCMC se producirá un incremento considerable del coste computacional. Se pueden encontrar algoritmos alternativos de complejidad equivalente o superior en Sengupta (1989) o Neuts (1981). Más recientemente, se ha obtenido que la distribución estacionaria del tiempo de espera en el sistema GI/PH/c, con c servidores, es también de tipo PH, véase Asmussen y Moller (2001). En este caso, el orden de la distribución de W se incrementa no sólo con el orden de la distribución del tiempo de servicio, sino también con el número de servidores, c, y por tanto, la dificultad computacional para incorporar estos resultados en un método de estimación será aún mayor.

Además del tiempo de espera, existe una literatura muy extensa sobre la construcción de algoritmos para obtener las distribuciones estacionarias de otras características asociadas a diferentes sistemas en los que intervienen distribuciones de tipo PH. Muchos de estos estudios se engloban dentro de los métodos matriciales introducidos por Neuts (1981). Por ejemplo, en Squillante (1998), se construyen algoritmos para obtener soluciones explícitas en forma matricial de la distribución estacionaria del número de clientes en cola en un sistema PH/PH/c. También, existen otros resultados, en el contexto matricial de Neuts (1981), en los que se analizan sistemas con llegadas que no son independientes, véase, por ejemplo, Ramaswami y Lucantoni (1985), o sistemas con interferencia en el acceso al servicio, véase, por ejemplo, Lillo y Neuts (1999). El problema con la metodología matricial es que, en algunos casos, especialmente si se trata de sistemas con más de un servidor, es necesario resolver ecuaciones matriciales no lineales de dimensiones

muy elevadas para obtener las distribuciones deseadas. Consecuentemente, se dificulta la combinación de estos algoritmos con los métodos MCMC puesto que se necesita resolver una de estas ecuaciones para cada observación de la muestra MCMC. Alternativamente a los métodos matriciales, existen procedimientos para obtener expresiones compactas de medidas asociadas a algunos sistemas de colas. Por ejemplo, Bertsimas (1990) propone un algoritmo para obtener, entre otras medidas, la distribución estacionaria del tamaño del sistema de colas MGE/MGE/c. Una metodología similar a ésta se utilizará en el Capítulo 5 para estimar el comportamiento transitorio del sistema MGE/MGE/1.

Por último, como se comentó anteriormente, las distribuciones de tipo PH tienen aplicaciones en otras áreas diferentes, aunque relacionadas con la Teoría de Colas, como son el Análisis de la Supervivencia, véase Huzurbazar (1999), o la Teoría del Riesgo, véase Asmussen (2000). Sería interesante estudiar si se pueden aplicar los métodos de inferencia propuestos en este Capítulo para estimar las cantidades de interés propias de estas materias, como por ejemplo, la probabilidad de ruina en procesos de riesgo, véase Bladt et al. (2003).